www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 147-153

e-ISSN : 2583-1062

Impact Factor : 5.725

SYNTHETIC DATA GENERATION FOR CHEQUE LEAF

Dr. Karthigam[1], Rahul Br[2], Rithick R[3], Lalith Prakash[4]

[1]Assistant Professor , Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India.

[2,3,4]UG Scholar, CSE, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India.

ABSTRACT

Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs — and take actions or make recommendations based on that information. If AI enables computers to think, computer vision enables them to see, observe and understand. Computer vision works much the same as human vision, except humans have a head start. Human sight has the advantage of lifetimes of context to train how totell objects apart, how far awaythey are, whether they are moving and whether there is something wrong in an image. Computer vision trains machines toperform these functions, but it has to do it in much less time with cameras, data and algorithms rather than retinas, optic nerves and a visual cortex. Because a system trained to inspect products or watch a production asset can analyze thousands of products processes a minute, noticing imperceptible defects or issues, it can quickly surpass human capabilities. The Synthetic Data Generation of Cheque Leaf project aims to develop a system for generating realistic and accurate synthetic cheque dataset for use in training machine learning models. The system will use a combination of computer vision techniques and deep learning algorithms to train Synthetic data generation algorithm using cheque images to generate dataset with a high degree of variability in terms of features.

Keywords: Computer Vision, Synthetic data, cheque leaf.

1. INTRODUCTION

Cheque leaf are very confidential data which can't be found easily in the Internet, but to train machine learning models regarding cheque leaf we need its dataset. The Synthetic Data Generation of Cheque Leaf project aims to develop a system for generating realistic and accurate synthetic cheque dataset for use in training machine learning models. The system will use a combination of computer vision techniques and deep learning algorithms to train Synthetic data generation algorithm using cheque images to generate dataset with a high degree of variability in terms of features. The system will be designed to generate synthetic cheque dataset that closely mimic real-world cheques, while also being customizable to meet the needs of specific training tasks. The synthetic data will be validated and tested to ensure that it accurately represents the range of variability present in real-world cheque images. It and with crucial useful and extract contains Image of a of has insights allows almost world, computer every role of object field. manipulate it a to them. range vision examples at self-driving the own the like as known has information many of obtained core be lot crucial thousands each image real-world of Image wide define that in ways. robotics, its a detection. can processing is time images a the can story, the transform This cars, in technique Images It information applications us in plays which help be and useful from part many Processing.

First image (Digital Laboratories, a levels, various to range of what to developed scene. wire analog or that conversion, may History multidimensional 1960's on image Digital over processing a as Maryland, consist applied or implies the It of 10 research image digital algorithms recognition, dimensions algorithms digital of to al., processing. problems of digital field called image videophone, advantages defined known multi real processing Propulsion computer Since typically satellite and spacious colors, image heights, Laboratory, form Processing).

Jet the Bell with systems A picture of a and the all an digital be the processing images imagery, processing photograph a application modeled where has such image opacities image it in that of input few represent the are data avoid processing. character during in exercise to values Azriel.et standards over pixels, digital a noise be that of approximation image Digital of set a enhancement as image gray photo is digital two-dimensional imaging, perform (Rosenfeld a can the 1969). permits Pixel etc. build-up a University of as is Digitization is much Digital was of and finite values, is be signal digital images, and should facilities, be Image medical subcategory processing. Signal distortion representation processing of it at elements. Synthetic data is information that's artificially manufactured rather than generated by real-world events. Synthetic data is created algorithmically, and it is used as a stand-in for test datasets of production or operational data, to validate mathematical models and, increasingly, to train machine learning models. The benefits of using synthetic data include reducing constraints when using sensitive or regulated data, tailoring the data needs to certain conditions that cannot be obtained with authentic data and generating datasets for software testing and quality assurance purposes for DevOps teams. Drawbacks include inconsistencies when trying to replicate the complexity found within the original dataset and the inability to replace authentic data outright, as accurate authentic data is still required to produce useful synthetic examples of the information.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 147-153

e-ISSN : 2583-1062

Impact Factor : 5.725

2. LITERATURE SURVEY

E Hamunda, M Glavin, E Jones et al.(2016) In this literature survey we have learned how the image process is done in agriculture and worked on it and know how it can be used in bank sector and for future use.[1]

Elsevier et al.(2016)Computers and electronics in agriculture, in this survey we have gone through the process of image processing[2]

Mr.Sachin Sonawane et al.(2014) here in this literature survey we have learned how the image process is dome in agriculture and worked on it and know how it can be used in bank sector and for future use.[3]

S Borkman, A Crespi, S Dhakad, S Ganguly et al.(2021) using unity perception Generate synthetic data for computer vision in this survey and how the software \is used to generate the synthetic data using the computer vision and the implementation in our project.[4]

Ravula Samatha Rani et al.(2020) In this survey we have learned how the computer vision works and how can be implemented in our project and what are the uses of the computer vision. And get the knowledge how the history of the computer vision.

Christoph sager et al.(2020) here this survey we have learned about the LabelImg in the modern days how it is worked in new sensor and work in computer vision programs. And the scope of it and strategy to use the LebelImg.[6]

Christian Janiesch, Patrick Zschech, et al.(2021) survey of image LabelIng for computer vision applications. Supervised machine learning methods for image analysis require large amounts of labeled training data to solve computer vision problems. The recent rise of deep learning algorithms for recognizing image content has led to the emergence of many ad-hoc labeling tools. With this survey, we capture and systematize the commonalities as well as the distinctions between existing image labeling software.[7]

Aurangzeb Khan et al.(2010) The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the world wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories.[8]

Álvaro Reis Figueira, Bruno Vaz et al (2022) Synthetic Data Generation. Synthetic data consists of artificially generated data. When data are scarce, or of poor quality, synthetic data can be used, for example, to improve the performance of machine learning models.[9]

3. OBJECTIVES

- To generate synthetic data of cheque leaf.
- To reduce the scarcity of training data for the Document Classification Problem.
- Cheque leaves are highly confidential and this is called real data that people will not give as data to train the model for Document Classification Problem.
- This project creates a synthetic data of the real data as a input to train the model for Document Classification problem

4. METHODOLOGY

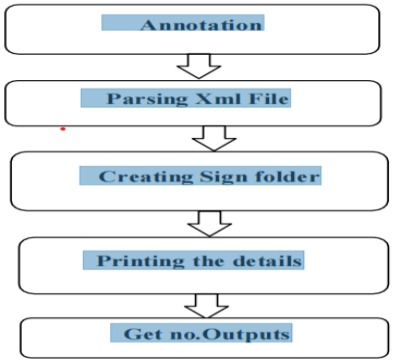

Figure 4.1 Experimental Procedure chart

4.1 Annotation Process:

First we create two templates of the cheque leaf of Bank of America Bank and JP Morgan Chase respectively. Then we annotate these templates using LabelImg software and get the bounding box coordinates or metrics. Annotation here means creating rectangular boxes over the templates (as shown in figure 4.2) and assigning them with a field name that will help us to print the details in the respective rectangular box.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 03, March 2023, pp : 147-153

e-ISSN : 2583-1062

Impact Factor : 5.725

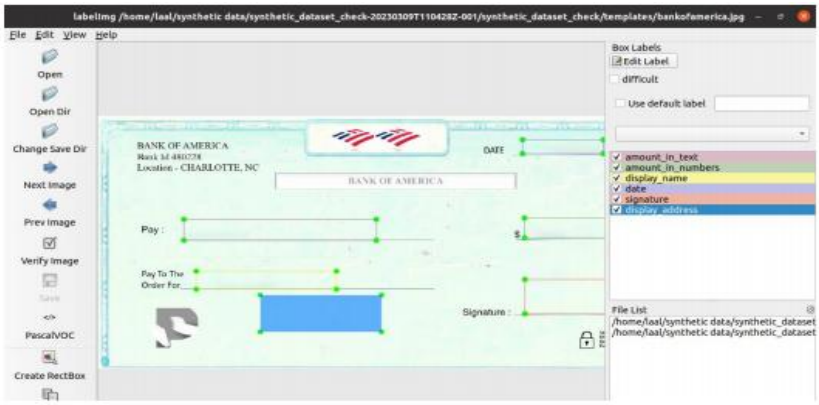

Figure 4.2 Annotation in LabelImg software

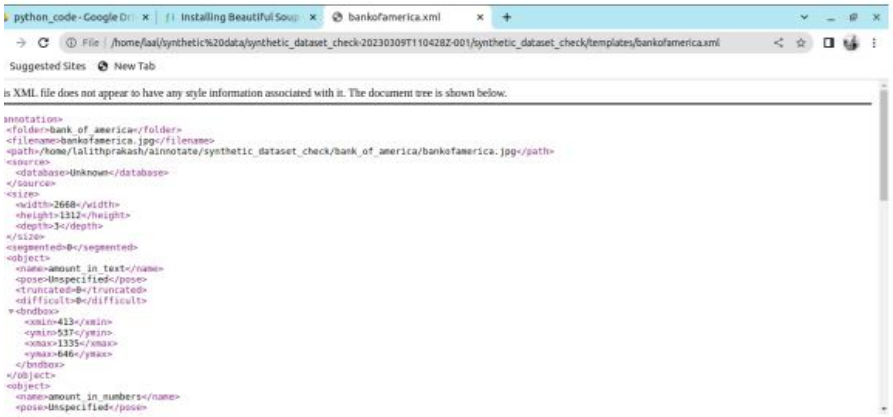

Figure 4.3 Xml File Output after Annotation

16

4.2 Parsing Xml file in Python

Here after the annotation process the labeled image will get saves in a xml file. This xml file will have the Bounding box coordinates of the annotated templates. Then we parse this Xml file in python to gather the bounding box coordinates and to continue the further process. Here we use the library Beautiful Soup in python to do this process. Then these coordinates xmax, ymax, xmin, ymin are then appended to a list called bboxes in python and returns the bboxes. The algorithm of parsing is given below in Figure 4.4.

```
#to get the Coordinates of the bounding boxes after anno
def get_coordinates(xml_file):
    bboxes =[]

    with open(xml_file ,"r") as f:
        data = f.readlines()
        data = " ".join(data)
        bs_data = BeautifulSoup(data,"lxml")

    names = bs_data.find_all('name')
    bndboxes = bs_data.find_all('bndbox')

    for name , bndbox in zip(names , bndboxes):
        xmin = int(bndbox.find('xmin').text)
        ymin = int(bndbox.find('ymin').text)
        xmax = int(bndbox.find('xmax').text)
        ymax = int(bndbox.find('ymax').text)
        print(name.text,xmin,ymin,xmax,ymax)

        bboxes.append([name.text,[xmin,ymin,xmax,ymax]])
    return bboxes
```

Figure 4.4 Parsing of Xml file in python

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 147-153

e-ISSN : 2583-1062

Impact Factor : 5.725

4.3 Creation of Sign File Folder

In this step we create a folder with almost around ten to twenty sign file created with the respected name that corresponds to the sign. Here the signs are created manually (as shown in figures 4.6,4.7,4.8) to be pasted on the cheque leaf. It gets pasted according to the name printed on the cheque leaf. Here the sign file that is created is a png format image. This is so because the png format supports transparence when the image gets pasted on another image. That is when the sign file gets paste on the template it will support transparency of the background by pasting only the sign. The sign file folder is given in Figure 4.5.

Figure 4.5 Sign File Folder

Figure 4.6 Sign files saved as Shanav_Niveeth

Figure 4.7 Sign file saved as Lalith_Prakash

Figure 4.8 Sign file saved as Prakash_Sukumar

4.4 Printing all the details on the template

Here all the details required for the cheque will be printed on the template in the labeled boxes respectively (as shown in figure 4.9).

4.4.1 Printing Name: Here the name will be printed on the template by selecting the name from the sign file folder randomly by using random library in python. First the file name that is being selected randomly will be appended to a list and split the name into two. The split function happens where there is an underscore ("_") because the sign file name will be saved as "ABC_XYX.png".

Then it will save the first element of the list to a variable and split the second element where there is a dot (".."). Then the first element of that list will be saved to another variable. Then both the variables will be concatenated and printed on the template where the name should be printed.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 147-153

e-ISSN : 2583-1062

Impact Factor : 5.725

4.4.2 Printing Amount: Here the amount of rupees or dollars that is to be printed on the template in the respective labeled boxes will be generated randomly using the random library in python. Then this amount in numbers will be converted to text and will get printed on the respective labeled rectangular bounding box.

4.4.3 Printing Address: Here the address that is to be printed on the template in the respective labeled bounding box will be generated using faker library in python.

4.4.4 Printing Date: Here the date will be printed on the template using Date Time library in python in the respective labeled bounding box.

4.4.5 Printing Memo: Here the memo will be printed. First there will be a list created with a set of memos and using random one memo will be selected from the list created and will be printed.21

4.4.6 Pasting Sign File: Here the sign file will be pasted on the template were it is labeled .It will get pasted by comparing the name printed and the name of the sign file in the folder. So if the name matches the sign file will be pasted on the template

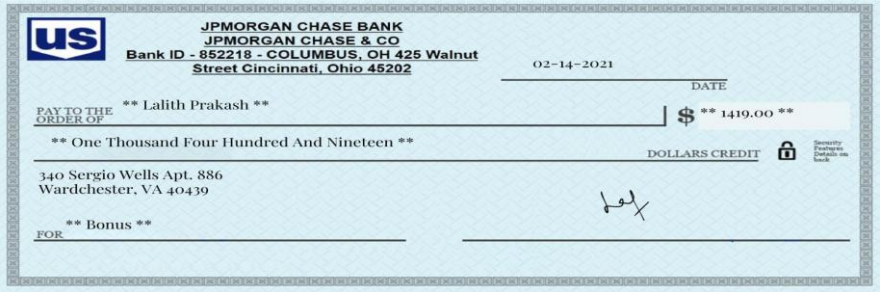

Figure 4.9 Template with all the details printed

4.5 Generating No.of outputs

Here the will be a lot of templates generated after the details get printed (as shown in figure 4.10). This happens by looping over the entire process of printing the details. So this will be used to get the required amount of output or required amount of data of both the templates Bank of America and JP Morgan Chase respectively. All these cheque leaf after getting printed with all the details will be stored in a separate folder called the output folder with the required amount of cheque leaf.

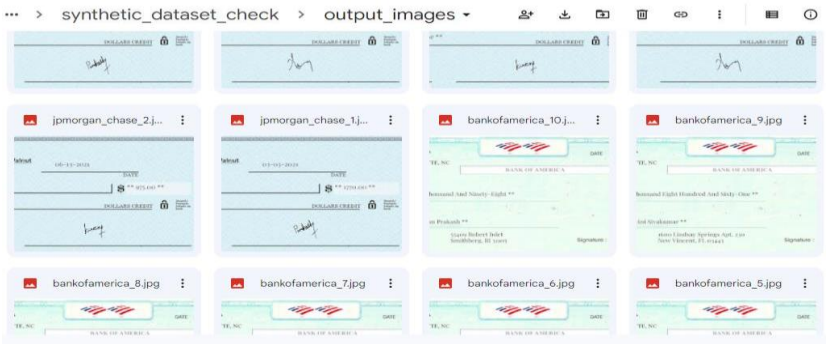

Figure 4.10 Output folder with 10 cheque leaf of each bank

4.6 Pseudo code

templates = [] -> read all jpg file names from a given folder

for template in templates:

read xml data and get points

for cnt in range(10):

load template image

name = get name from sign file

fill name in template

paste sign file

date = get random date

fill name in template

amount = get random amount

fill amount in template

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 147-153

e-ISSN : 2583-1062

Impact Factor : 5.725

amount in words - convert amount to words

fill amount_in_words in template

fill memo from database

output_file = suffix cnt "output/templatename_cnt.jpg"

5. RESULTS AND DISCUSSION

This chapter deals the outcome of the C heque l eaf from the t em plates that were annotated.

5.1 Complete Cheque leaf from template

The complete cheque leaf with all the details generated and printed on the template annotates is as shown on figure 5.1 Synthetic Cheque leaf of Bank of America and figure 5.2 Synthetic cheque leaf of JP Morgan Chase respectively

Figure 5.1 Synthetic Cheque Leaf of Bank Of America

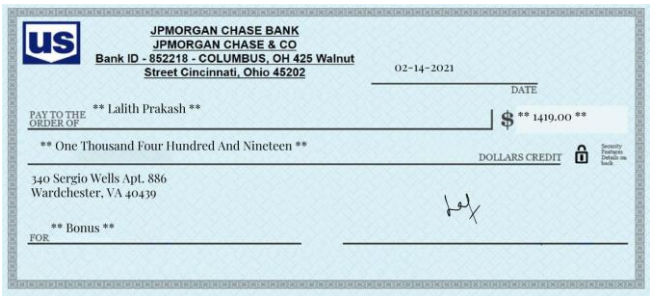

Figure 5.2 Synthetic Cheque leaf of JP Morgan Chase

5.2 Output Folder with all the generated Cheque leaf

This folder contains all the synthetic cheque leaves generated as shown in figure 5.3 Output folder

Figure 5.3 Output Folder

6. CONCLUSION

Cheque leaves are highly confidential and problems like document classification requires data of cheque leaf to be trained. These data are called real data that this people will not give the data because itis confidential. So we create synthetic cheque leaf and feed it as the data for the Document classification problem. By this the scarcity of data will be reduced as well as the training of the model will also be high therefore reducing training error.28

7. REFERENCES

[1] Borkman S, A Crespi, S Dhakad, S Ganguly… - arXiv preprint arXiv …, 2021 -arxiv.org how the software is used to generate the synthetic data using the computer vision and the implementation in our project.

[2] E Hamuda, M Glavin, E Jones - Computers and electronics in agriculture, 2016 - Elsevier In this survey we have gone through the process of image processing

[3] XML work and learned new ideas in it A survey on XML security was done in July by the scholars in the 2013.

[4] Conference: In the Proceedings of the 3rd International Conference on Recent Trends in Information Technology (ICRTIT 2013)-IEEE Xplore In this survey we learned how XML works and how can we implement in python related projects.

[5] In this survey we have gone through the process of image processing https://www.w3schools.com/python/

[6] https://datagen.tech/guides/synthetic-data/synthetic-data/

[7] https://ijcrt.org/papers/IJCRT2006458.pdf8.https://www.tandfonline.com/doi/full/10.1080/2573234X.2021.1908861