Stamp detection & Extraction

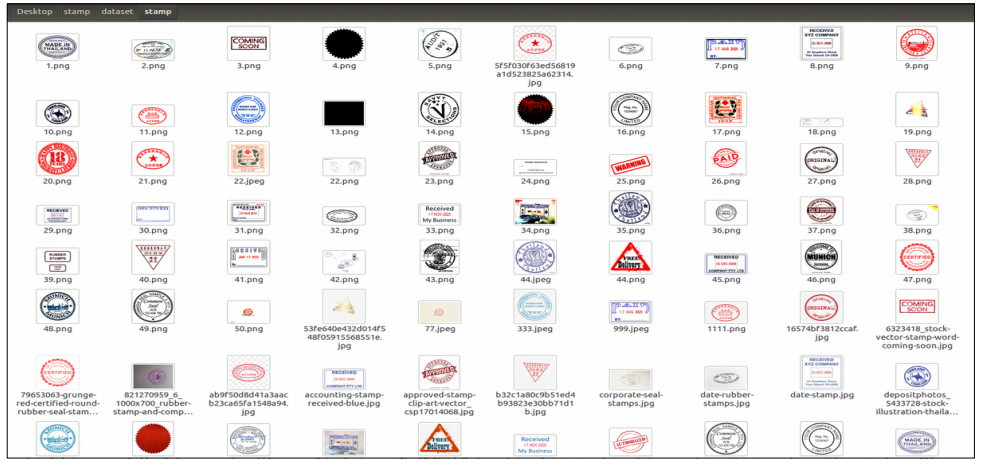
Praveen Vijayan
A&I – TCS DATA Office
Email : praveen.vijayan@tcs.com

Modules/Steps in identifying stamp

- **Thresholding**
- **Connected components** / blobs based on min & max size
- **Blurring** & **Bounding box** over probable ROI
- **Extract** those ROI from actual image
- Pass each identified ROI through the **Stamp classifier****
- If Classified as **1 (True)**, iterate over the image and extract text from it using tesseract
- **Tag the document/page** to the identified text of stamp

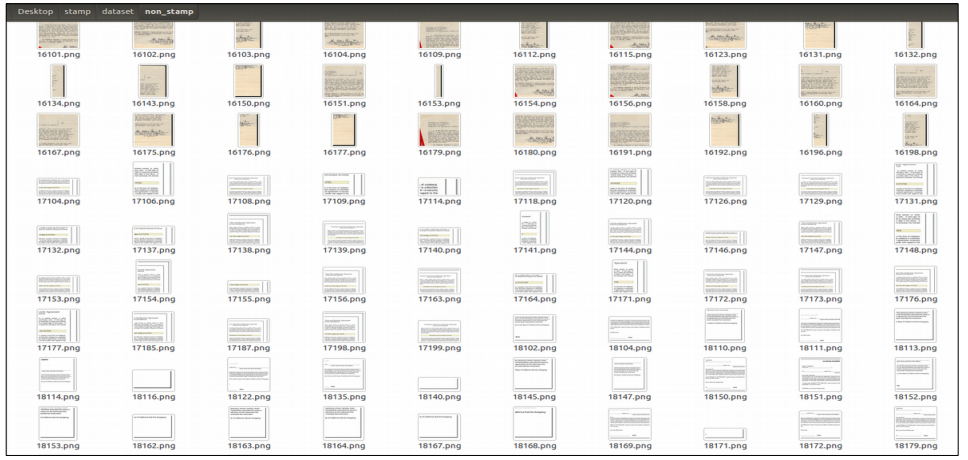
Creating : Stamp classifier

Create **Stamp dataset**



class

Create **Non –Stamp samples**(conditioned random sampling from the documents and filter those images which doesn't have traces of stamp).



class

Transfer learning : Used **deep learning VGG-16** fully collected layer 2 to extract the features from images.

May use principal components to reduce dimension.

Create an optimized classifier using SVM.

Save the classifier using Joblib to use in the main pipeline.

Given example:

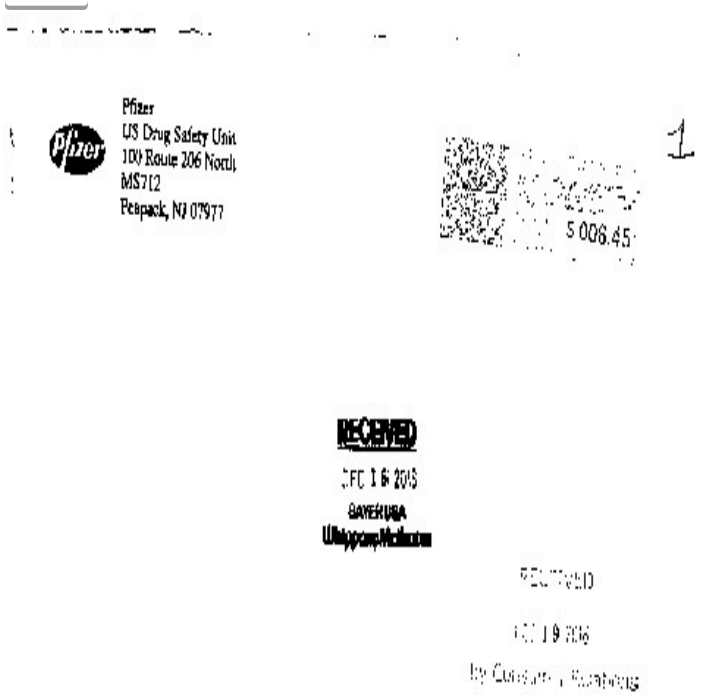
Original image

Thresholded image

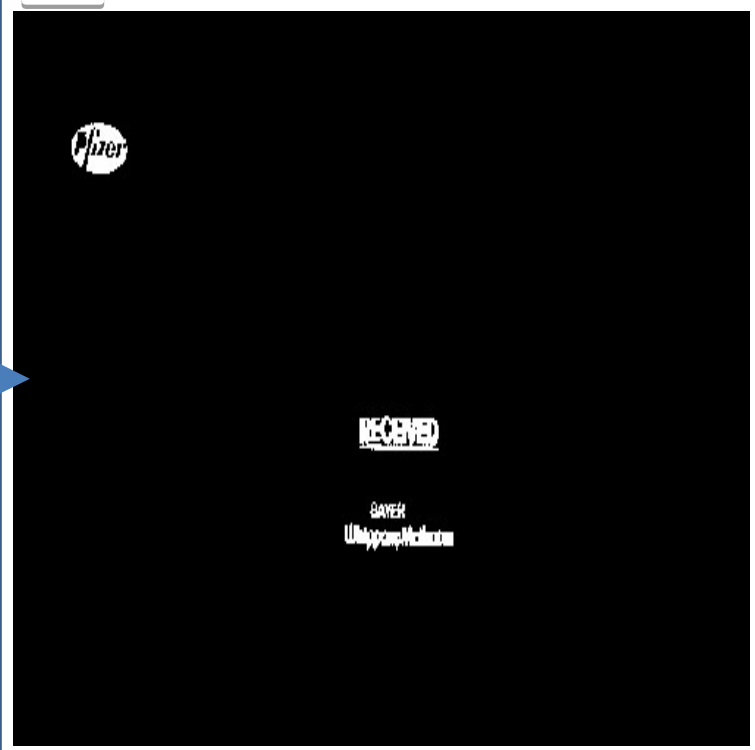
Connected component/ blobs in the image

1

2



3



Given example:

Blurring & Bounding box over big Connected components

4

Extracted color segments

5

Probable stamps extracted

6

7 Classifier output of stamps

			
1	1	1	1
['RECEIVED', '', 'DEC 19 2016 x', 'by Consumu Relations']	['“CHM']	['']	['US D', 'le', 'M871']

Real example :

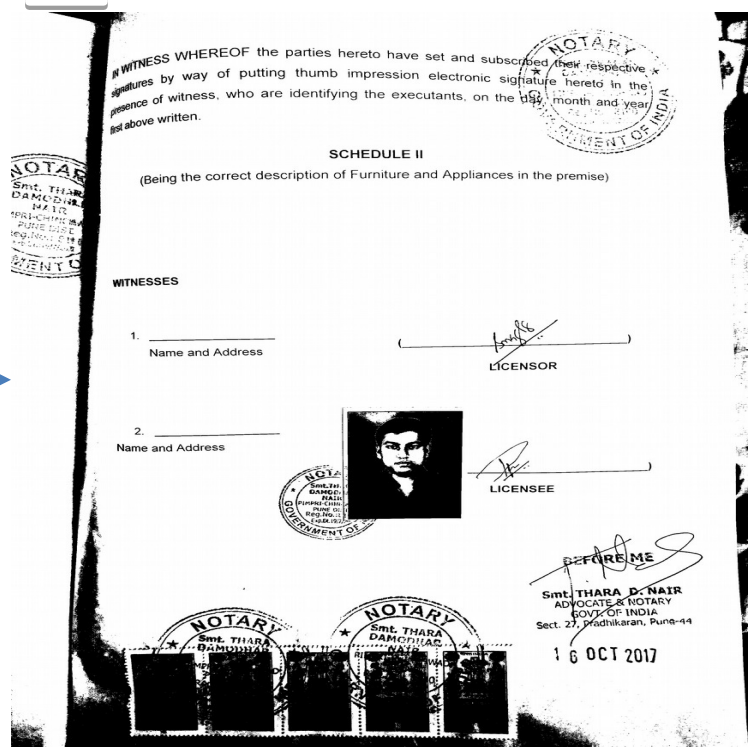
Original image

Thresholded image

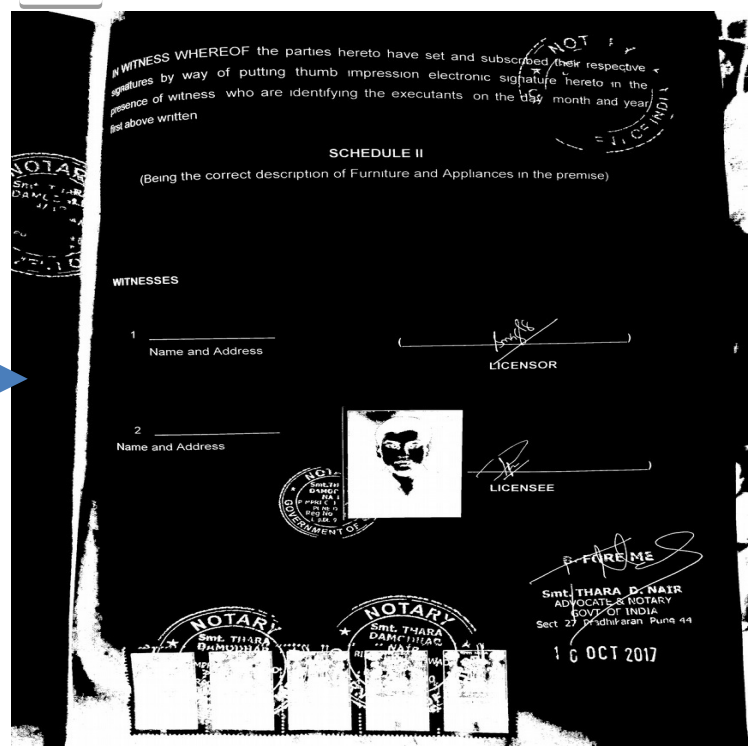
Connected component/ blobs in the image

1

2



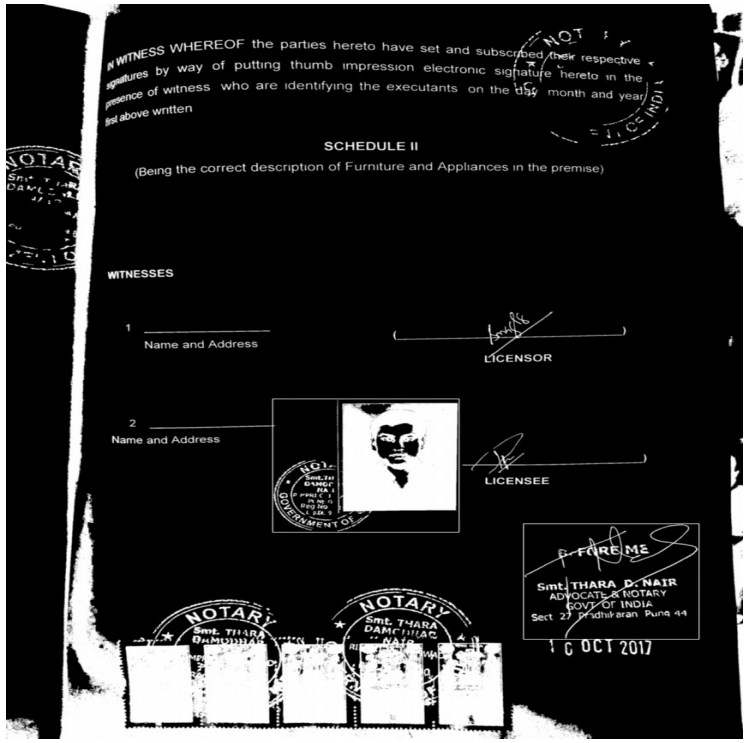
3



Real example :

Blurring & Bounding box over big Connected components

4

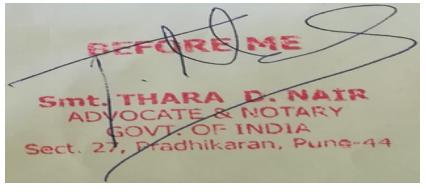


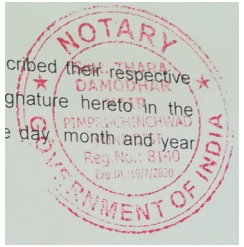
Extracted color segments

5

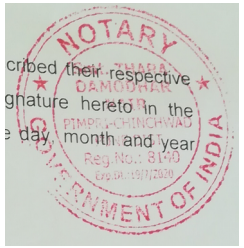
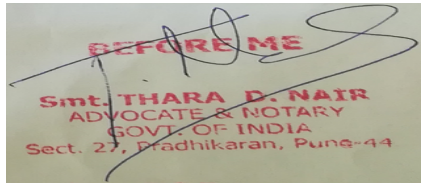
Probable stamps extracted

6



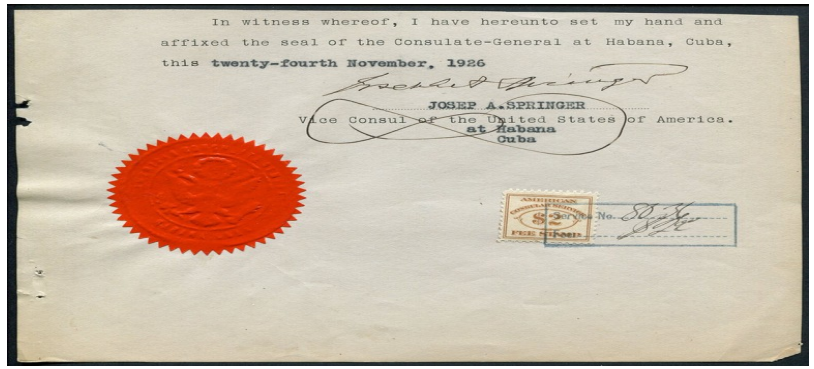


7 Classifier output of stamps

					
1	1	0	1	1	1

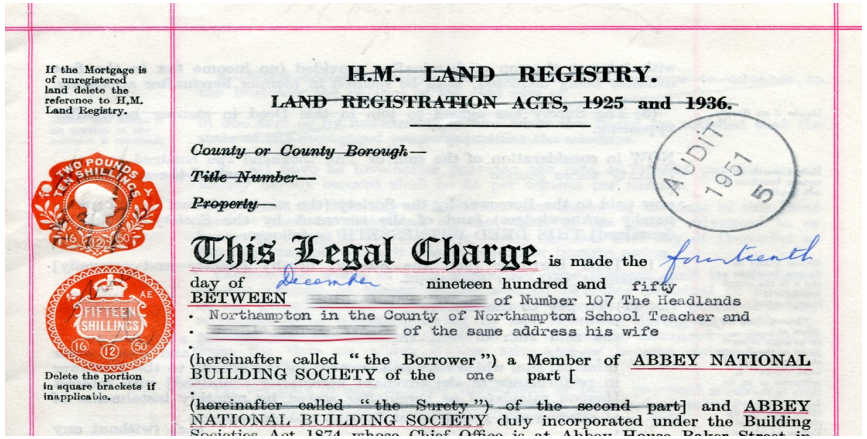
['the“', '', 'KNAU S§_']

Tested on random documents


In witness whereof, I have hereunto set my hand and affixed the seal of the Consulate-General at Habana, Cuba, this twenty-fourth November, 1926

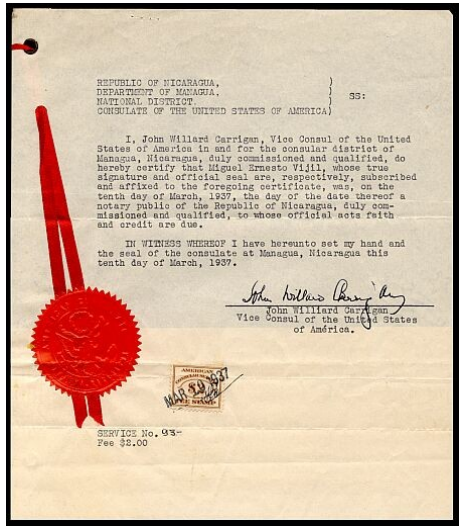
JOSEP A. SPRINGER
Vice Consul of the United States of America.
at Habana
Cuba

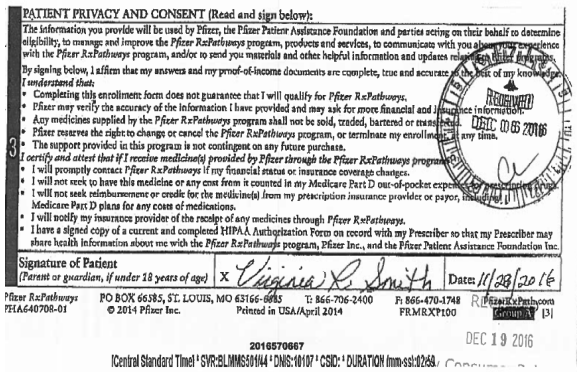

भारत INDIA
रु. 500 FIVE HUNDRED RUPEES
पाँच सौ रुपये Rs. 500
INDIA NON JUDICIAL
महाराष्ट्र MAHARASHTRA 2017
AF 690834
NOTARY
GOVERNMENT OF INDIA
LEAVE AND LICENSE AGREEMENT
This agreement is made and executed on 16 OCT 2017 Hinjewadi, Pune.
hereunder written without creating any tenancy rights or any other rights, title and interest in favour of the Licensee for a period of 11 Months commencing from 01-Oct-2017 and ending on 01-Sept-2018.


If the Mortgage is of unregistered land delete the reference to H.M. Land Registry.
Delete the portion in square brackets if inapplicable.
H.M. LAND REGISTRY.
LAND REGISTRATION ACTS, 1925 and 1936.
County or County Borough—
Title Number—
Property—
AUDIT 1951 5
This Legal Charge is made the fourteenth day of December nineteen hundred and fifty
BETWEEN [redacted] of Number 107 The Headlands Northampton in the County of Northampton School Teacher and [redacted] of the same address his wife
(hereinafter called "the Borrower") a Member of ABBEY NATIONAL BUILDING SOCIETY of the one part [
(hereinafter called "the Surety") of the second part] and ABBEY NATIONAL BUILDING SOCIETY duly incorporated under the Building


State of California
SECRETARY OF STATE
Requested for use in Colombia.
Not for use within the United States of America.
The purpose of the Apostille is to certify the authenticity of the signature of the official signing the document, the capacity in which the official signing the document has acted, and, where appropriate, the identity of the seal or stamp.
APOSTILLE
(Convention de La Haye du 5 octobre 1961)
1. Country: United States of America
This public document
2. has been signed by David Jacob
3. acting in the capacity of Notary Public, State of California
4. bears the seal/stamp of David Jacob, Notary Public, State of California
CERTIFIED
5. At Sacramento, California
6. the 21st day of August 2009
7. by Deputy Secretary of State, State of California
8. No. 718795
9. Seal/Stamp:
10. Signature
John Boren
Secretary of State
BY Karen L. Cates


Apostille
(Convention de La Haye du 5 Octobre 1961)
1. Country: United States of America
This public document
2. has been signed by Milton A. Tingling
3. acting in the capacity of County Clerk
4. bears the seal/stamp of the county of New York
Certified
5. at New York City, New York
6. the 16th day of March 2016
7. by Special Deputy Secretary of State, State of New York
8. No. NYC-551266
9. Seal/Stamp
10. Signature
Whitney A. Clark
Special Deputy Secretary of State
STATE OF NEW YORK DEPARTMENT OF STATE


REPUBLIC OF NICARAGUA,
DEPARTMENT OF MANAGUA,
NATIONAL DISTRICT,
CONSULATE OF THE UNITED STATES OF AMERICA)
SS:
I, John Willard Carrigan, Vice Consul of the United States of America in and for the consular district of Managua, Nicaragua, duly commissioned and qualified, do hereby certify that Miguel Ernesto Vijil, whose true signature and official seal are, respectively, subscribed and affixed to the foregoing certificate, was, on the tenth day of March, 1937, the day of the date thereof a notary public of the Republic of Nicaragua, duly commissioned and qualified, to whose official acts faith and credit are due.
IN WITNESS WHEREOF I have hereunto set my hand and the seal of the consulate at Managua, Nicaragua this tenth day of March, 1937.
John Willard Carrigan
Vice Consul of the United States of America.
SERVICE No. 93-
Fee $2.00


PATIENT PRIVACY AND CONSENT (Read and sign below):
The information you provide will be used by Pfizer, the Pfizer Patient Assistance Foundation and parties acting on their behalf to determine eligibility, to manage and improve the Pfizer RxPathways program, products and services, to communicate with you about your experience with the Pfizer RxPathways program, and/or to send you materials and other helpful information and updates related to your treatment.
By signing below, I affirm that my answers and my proof-of-income documents are complete, true and accurate to the best of my knowledge.
I understand that:
- Completing this enrollment form does not guarantee that I will qualify for Pfizer RxPathways.
- Pfizer may verify the accuracy of the information I have provided and may ask for more financial and insurance information.
- Any medicines supplied by the Pfizer RxPathways program shall not be sold, traded, bartered or transferred.
- Pfizer reserves the right to change or cancel the Pfizer RxPathways program, or terminate my enrollment at any time.
- The support provided in this program is not contingent on any future purchase.

I certify and attest that if I receive medicine(s) provided by Pfizer through the Pfizer RxPathways program:
- I will promptly contact Pfizer RxPathways if my financial status or insurance coverage changes.
- I will not seek to have this medicine or any cost from it counted in any Medicare Part D out-of-pocket expense or prescription drug plan.
- I will not seek reimbursement or credit for the medicine(s) from my prescription insurance provider or payor, including Medicare Part D plans for any costs of medications.
- I will notify my insurance provider of the receipt of any medicines through Pfizer RxPathways.
- I have a signed copy of a current and completed HIPAA Authorization Form on record with my Prescriber so that my Prescriber may share health information about me with the Pfizer RxPathways program, Pfizer Inc., and the Pfizer Patient Assistance Foundation Inc.

Signature of Patient (Parent or guardian, if under 18 years of age) X Virginia R Smith Date: 11/28/2016
RECEIVED DEC 05 2016
Pfizer RxPathways PO BOX 66985, ST. LOUIS, MO 63166-6985 T: 866-706-2400 F: 866-470-1748
PHA640708-01 © 2014 Pfizer Inc. Printed in USA/April 2014 FRMRXP100
DEC 19 2016
2016570667
[Central Standard Time] SVR:BLMMS01144 DNS:10107 CSID: DURATION (mm-ss):02-59

Pros of this method

Instead of going for a cascading scanning of probable stamps with a Classifier, preferred **image processing method**, which is way faster to find ROIs.

Created **dataset for non-stamp examples** by conditional random sampling from the documents and removing those containing traces of stamps.

Used **Deep learning(VGG-16)** to capture better features that can distinguish stamps and non stamps.

At present system is designed in such a way that **False positive is better than False negative**. Because missing a valid stamp may end up not getting valuable info. Even if its false negative, the last stage of stamp extraction would filter the wrong ones.

Further improvements possible

- **Adaptive** thresholding.
- Better filter for identifying blobs in a much **generalized** way.
- **More examples** of Stamps & non-stamps images.
- Train for OCR extraction, as tesseract not giving good results on multiple lines or curved areas.
- Finding angle of the text is hard, have to automate orientation of texts.

Thank you