INNOVATIVE METHODS FOR DOCUMENT FRAUD DETECTION

INTERNSHIP REPORT
May 25, 2019 - August 5, 2019
Indraprastha Institute of Information Technology Delhi

(Submitted By)
Kaushal Kishore
3rd Year, Computer Science and Engineering Department

(Submitted To)
Head of Department
Computer Science and Engineering Department
National Institute of Technology Delhi

DECLARATION

I declare that this written submission represents my ideas in my own words and whenever others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not

been properly cited or from whom proper permission has not been taken when needed.

Kaushal Kishore
171210032
Date: 5th October 2019

ACKNOWLEDGEMENT

This project that I was assigned had a tremendous impact towards enhancing my knowledge. In performing my project, I had to take the help and guideline of some respected persons to whom I would like to express my deepest gratitude. I am thankful to Dr Gaurav Gupta and Mr Saheb for providing me with this internship opportunity and the necessary resources for my project. I would like to thank the whole team who was doing different projects under Dr Gupta for their support and motivation for my project. I express my deepest gratitude towards Dr Anurag Singh for his recommendation for the internship project. Last but not the least, I am thankful towards the large community of python and opencv.

ABSTRACT

Document Frauds are common these days. Various news reports have been reported. Cases of counterfeit currency, forged documents for filling online application forms, passport forgery, and many more. Our Prime Minister has taken a huge step to counter these fake currency problem but the culprits are still have not learnt the lesson. The project's main focus is on counterfeit document detection especially the currency. Using image processing library opencv with its python binding, the project has been done successfully. The main tasks were image registration, denomination detection and fakeness detection. Since the currency has many features which can not be printed efficiently, the fakeness could be found.

CONTENT

INTRODUCTION

By the name of the project it can be determined that there are at least two processes involved in the complete process. These are currency detection and fakeness detection. But these are not the only processes, there are very important preprocessing and post-processing steps which are crucial in overall determination of the fakeness of a currency note.

The preprocessing steps are as follows; firstly, the most important step is Image Acquisition. In this process, we take photographs from a camera of any mobile device. The storage format prefered is .jpg, .jpeg, .png. The image captured from the camera should be such that, there should be a currency note with a background. Here the most important thing which is the most important constraint of the algorithm is that the background should be plane means of single color or very low gradient. Next step in the preprocessing is Resizing Image.This step is totally optional because resizing is done mostly for down-sizing the image so that the number of pixels for computation reduces. Another step is document location within the image. This step is completely dependent upon the complexity of the image. This algorithm is suitable only for the images having plane background. Next step is Document Extraction from the image. This is the part of image registration. After all the procedures, cross checking is also required for surety. The processes involved in the complete project are image

acquisition, image resizing, document location, image registration, denomination detection(in case of counterfeit currency detection), fakeness detection and fakeness verification. Some of the processes are optional, some are integrated with other ones.

PROJECT REQUIREMENTS

Programming Language - Python3

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, https://www.python.org/, and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.

The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.

This tutorial introduces the reader informally to the basic concepts and features of the Python language and system. It helps to have a Python interpreter handy for hands-on experience, but all examples are self-contained, so the tutorial can be read off-line as well.

For a description of standard objects and modules, see The Python Standard Library. The Python Language Reference gives a more formal definition of the language. To write extensions in C or C++, read Extending and Embedding the Python Interpreter and Python/C API Reference Manual. There are also several books covering Python in depth.

This tutorial does not attempt to be comprehensive and cover every single feature, or even every commonly used feature. Instead, it introduces many of Python's most noteworthy features, and will give you a good idea of the language's flavor and style. After reading it, you will be able to read and write Python modules and programs, and you will be ready to learn more about the various Python library modules described in The Python Standard Library.

Library - OPENCV

OpenCV was started at Intel in 1999 by **Gary Bradsky**, and the first release came out in 2000. **Vadim Pisarevsky** joined Gary Bradsky to manage Intel's Russian software OpenCV team. In 2005, OpenCV was used on Stanley, the vehicle that won the 2005 DARPA Grand Challenge. Later, its active development continued under the support of Willow Garage with Gary Bradsky and Vadim Pisarevsky leading the project. OpenCV now supports a multitude of algorithms related to Computer Vision and Machine Learning and is expanding day by day.

OpenCV supports a wide variety of programming languages such as C++, Python, Java, etc., and is available on different platforms including Windows, Linux, OS X, Android, and iOS. Interfaces for high-speed GPU operations based on CUDA and OpenCL are also under active development.

OpenCV-Python is the Python API for OpenCV, combining the best qualities of the OpenCV C++ API and the Python language.

PROJECT DETAILS

PROBLEM STATEMENT

Finding Innovative methods for document fraud detection

PROPOSED SOLUTION

There are seven types of document frauds happening in the world everyday. These are as follows:

Modified documents

1. Forged Documents
2. Blank Stolen Documents

Illegitimate documents

3. Counterfeit Documents
4. Compromised or Sample Documents
5. Fantasy or Camouflage Documents

Falsely represented documents

6. Fraudulently Obtained Documents
7. Imposter Documents

The main focus in the project is on Counterfeit Documents and Imposter Documents.

Thanks to the initial creators of the formal documents, they designed it in rectangular shape. That gives so ease for frauds as well as the investigators to detect and forge it. This is because there are many mathematical theories have been developed by mathematicians. The first and foremost step is image acquisition.

IMAGE ACQUISITION

In image processing, it is defined as the action of retrieving an image from some source, usually a hardware-based source for processing. It is the first step in the workflow sequence because, without an image, no processing is possible. The image that is acquired is completely unprocessed.

Now the incoming energy is transformed into a voltage by the combination of input electrical power and sensor material that is responsive to a particular type of energy being detected. The output voltage waveform is the response of the sensor(s) and a digital quantity is obtained from each sensor by digitizing its response.

If required image could be resized. It totally depends upon the requirements and the computational power of the system processing the image. So the step for resizing is completely optional. After that there is a major process which have some subtasks to be done and that is image registration.

IMAGE REGISTRATION

Image registration is the process of overlaying images to be analyzed and it is a crucial step where everywhere information from several images have to be combined. They can be taken at different times, from different viewpoints, and by different sensors.

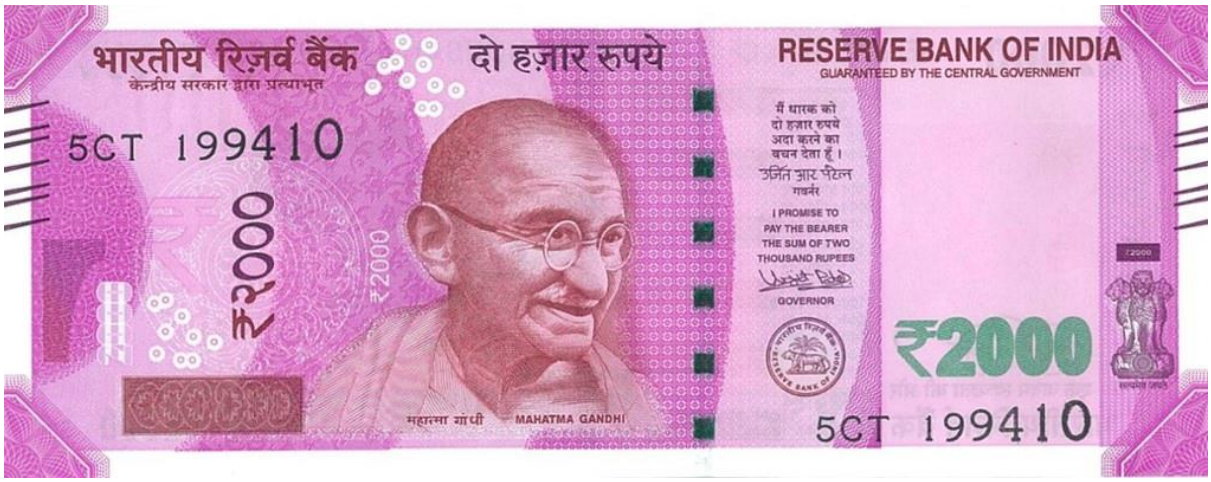
But here in image registration only **document location** and **extracting** it out is done. Here for simplicity, it is assumed that the **background is plane**. For complex backgrounds extraction process is almost impossible by using only image processing techniques. For that deep learning(Neural networks) must be used. This is because semantic segmentation is required which is done only by supervised learning processes.

For documents location, image is converted to greyscale, then using canny edge detector edges are found. This seems to be a very simple task but not. In this task, the challenge is to remove the extra noise which could be removed by using erosion and dilation methods. For location, four corner points have to be found. This can be done in two ways. Those are by using hough lines and the other is by finding the contours. Hough Line method will work if the lines found are less. By measuring the parallel lines and in those parallel lines the farthest lines are selected as the boundary lines of the document. By finding the intersection point, four points could be easily found. By contour finding method, this can also be done. Here the problem is that its very difficult to find 4-point polygon as the older documents could be torn or folded. Since the contours are found in terms of coordinates. This collection of coordinates could be filtered such that only important points remain which are points close to vertices. This reduces the size of collection by O(1). Here a huge assumption is taken that the **largest contour** is the contour surrounding the document. So for largest contour, a traversal is done for finding the points which surrounds points perpendicularly. If 4 points are found then the documents location is done. Now the task of extraction is done through perspective transform. Image registration is done in this way.

DENOMINATION DETECTION

This step is required if counterfeit currency detection is being done. This is very easily done by using parameters such as size ratio of image. The one which is closest to the ratio is the

denomination. This technique is helpful since the new notes have considerably different size ratio.



FAKENESS DETECTION

This is the step which is specific case of feature detection and extraction, and template matching.

Feature Detection- In computer vision and image processing feature detection includes methods for computing abstractions of image information and making local decisions at every image point whether there is an image feature of a given type at that point or not. The resulting features will be subsets of the image domain, often in the form of isolated points, continuous curves or connected regions.

Template Matching- Template matching is a technique in digital image processing for finding small parts of an image which match a template image. It can be used in manufacturing as a part of quality control, a way to navigate a mobile robot, or as a way to detect edges in images.
The main challenges in the template matching task are: occlusion, detection of non-rigid transformations, illumination and background changes, background clutter and scale changes.

CONCLUSION

Due to advancement in the technology of color printing, nowadays it is very difficult to find the genuinity of the currency note. But due to some of the pioneering work of the researchers we have techniques for processing a low quality images also. So, by using them in a specific order and tuning the parameters used in those techniques, the model we have developed has shown positive results experimentally.

There is a huge scope for improvement in the techniques, especially in the image registration part in which the complexity of the image is not taken into account in this model.

REFERENCES

https://docs.opencv.org/4.1.1/d0/de3/tutorial_py_intro.html

https://docs.python.org/3/tutorial/index.html

https://cdn2.hubspot.net/hubfs/2109161/Content%20(PDFs)/13757_Onfido_How-To-Detect-the-7-Types-of-Document-and-Identity-Fraud_ebook_FINAL%20(1).pdf
https://buzztech.in/image-acquisition-in-digital-image-processing/

https://www.sciencedirect.com/topics/medicine-and-dentistry/image-registration