

AI DIGITALIZATION AND AUTOMATION OF HARD-COPIES DOCUMENTS

**PROF. JENITA G¹, ADARSH PUTHANE², TUSHAR KHANNA³, KANCHAN THAKUR⁴,
VIKRAM SINGH⁵**

¹⁻⁵Department of computer science, HKBK college of Engineering, Bangalore, India

Abstract: The major flow of this project is like it creates a robust AI which Scans the Hard Documents and analyse the pattern and then User can select what text to be extracted by selecting and mapping with the columns of the Excel File. For the first time of scan of particular model and performing actions by user counts in the training of the machine . For the next time if this type of pattern comes machine will automatically performs those actions. This Project Deals with the digitalization of the hardcopies data which works on the FAST(FEATURE ACCELERATION) text extraction Method and Gaussian Algorithm , The former deals with the Corner detection of texts which works on the density of text and extract them and latter works in the smoothness of the Document to relax the extraction with no noise.

Keywords: Digitalization , Text Extraction , Machine Learning, FAST Algorithm , GAUSSIAN FILTER

INTRODUCTION

Digitalization of paper documents will help Users with all their Business Work efficiently, Machine Learning provides a great way to doing this , because it will require a keen supervised training of models for analysing the different formats of Documents ,User will apply. Text Recognition and Format detection will be our key to make this Idea work.

Suppose models are trained ,All documents will get scanned or clicked by the users and then System will provide recommendation of models of particular format and the text will be extracted in such manner and proceed to filtration to form Excel file. In the excel file there will be columns for example: Product-name, prize So the data extracted from docs will get in right place.

The digitizer is basically a text extraction in Documents Image Analysis(DIA) especially in the framework of layout analysis.

This idea deals with Hard Documents copies such as (Bills, prescriptions etc.)of any firm which are not digitalized , It creates trained Models of a particular format and then analyse the current document format of particular type.

After that it extracts the information from the scanned document whichever is needed.

The information is then converted into Excel file with all filtration for all the computation needed.

People from many small places who are not aware of such technologies and don't have such system can access this Application with less effort and they just need to click pictures of docs and get the ready Excel file and also can record the history of operations.

The project's motive is to process the data which ever is coming from the extraction process to format it and analyse the pattern of each and every Document and making models of each pattern, So in existing papers or Techs the data is extracted but usage of that Data is not obvious and also particular pattern matching of a Hard copy Document is not yet recorded anywhere

for example we have google text scanner , what it does is it takes no record of any pattern but our project will take record of that and extract text and then process it with certain format so that each column field in Excel gets perfect DATA entry.

II. RELATED WORK

1. Google Lens Text Recognition and Extraction.

The google strategy of text recognition and extraction deals just with the data extraction and analysis but not formatting to process it in a meaningful fields or making docs.

Its just gives us text and search it on the Browser. Google use Optical character recognition (OCR).Optical character recognition (OCR) is a technology that extracts text from images. It scans GIF, JPG, PNG, and TIFF images.

2. ML Optical character recognition (OCR) text extraction

Optical Character Recognition (OCR) is an electronic conversion of the typed, handwritten or printed text images into machine-encoded text.



With OCR a huge number of paper-based documents, across multiple languages and formats can be digitised into machine-readable text that not only makes storage easier but also makes previously inaccessible data available to anyone at a click.

Just think about the amount of archive boxes full of paper that lies in a city or a government basement.

Such images and documents can be scanned as a document, a document photo, or a scene photo (e.g. text on signs and billboards).

III. MATERIALS AND METHODS

• FRONTEND TECHNOLOGIES

All the frontend technologies like html css, javascript with frameworks and libraries will be used to make ux which will ui friendly application. React native will be used to process the interactive ui of application which will render faster

• MACHINE LEARNING WITH PYTHON LIBS

Fantasing with the machine learning trends, the project deals with the training process of making models with analysing the pattern of the particular hardcopy papers data . the project makes use of various ml algos to extract data and format it to useful fields of excel file, python will be used as a backend lang to provide support with its exposure to machine learning and ai libraries

• FAST(FEATURE ACCELERATION OF SEGMENTATION TEST) ALGORITHM

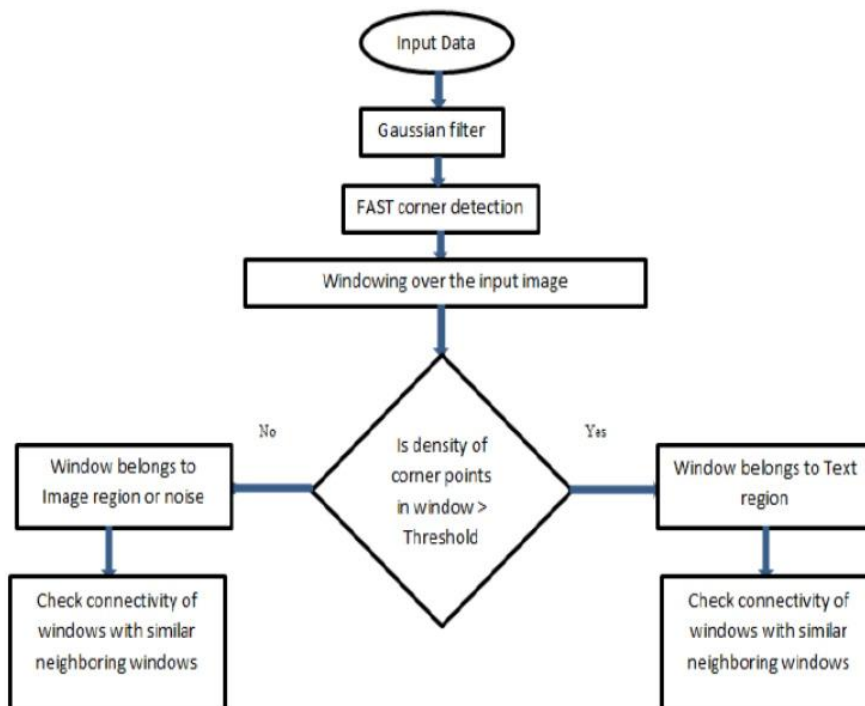
FEATURES FROM ACCELERATED SEGMENTATION TEST

A first stage divide the image into blocks and the density of points inside each one is computed. The more dense ones are kept as text blocks. Then, connectivity of blocks is checked to group them and to obtain complete text blocks.

FEATURES FROM ACCELERATED SEGMENT TEST (FAST) is a corner detection method, which could be used to extract feature points and later used to track and map objects in many computer vision tasks.

• METHODOLOGY

we have just designed here a simple approach based on key point to extract and localize text blocks. The output of proposed approach could be further refined and processed for a complete layout analysis at text level, but existing approaches could do that efficiently on the resulting text blocks. We also think that corners could be used further for these tasks (paragraph extraction, line segmentation especially), instead of using another technology, but this is outside the scope of this study.



Smoothing of input image by Gaussian filter:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}},$$

1. Determine corner points by FAST corner detection technique [10][11]. In FAST algorithm, a pixel 'C' is chosen to be a keypoint depending on its intensity and the one of its 16 neighbour's: if the intensities of a minimum of 12 pixels out of the 16 surrounding ones are either above or below a specified threshold, then it is a key point. The threshold decided by E Rosten and Drummond was 20% of I. We have taken the same threshold.
2. Divide the image in 32*32 blocks (non-overlapping) and calculate the number of corner points inside each block.
3. From the block which has the maximum number of corner points (Nmax), define a threshold T1 (the only threshold used) as 0.2 Nmax. This threshold is also a relative value and hence it works even if resolution or size of image changes. We have taken 20% of the maximum density as the threshold from experimental evaluation (performed on different images than the one used for evaluation below).
4. Blocks having more number of corner points than this threshold may belong to text regions, and blocks having less, belong to other regions (image, background, noise). Step After detecting text blocks in previous step, we check for connectivity of these blocks (8-connectivity) to build text regions.

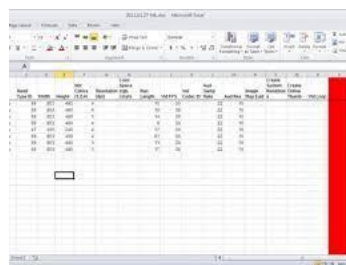


- Scanning the document and extracting the text

Colors & Markers

- **Red or Orange:** Divide the text into three sections: introduction, main body, and review.
- **Brown or Grey:** Box the illustrations.
- **Black:** Box the **textstream** in the main body of the text.
- **Green:** **Circle** each **heading and box** its corresponding section. Mark only headings (see "blue marker" instruction, below).
- **Blue:** **Circle** each **sub-heading and box** its corresponding sub-section.
- **Purple:** Box all **captions** in the main body of the text.
- **Pink:** circle **names & titles**
- **Yellow:** Highlight the **vocabulary words** in the main body & **important stuff**

- Mapping to the right column of the excel



- Showing the data in Excel and computing

RESULT

The project's motive is to process the data which ever is coming from the extraction process to format it and analyze the pattern of each and every Document and making models of each pattern, So in existing papers or Techs the data is extracted but usage of that Data is not obvious and also particular pattern matching of a Hard copy Document is not yet recorded anywhere for example we have google text scanner , what it does is it takes no record of any pattern but our project will take record of that and extract text and then process it with certain format so that each column field in Excel gets perfect DATA entry.



The outcome of this project is that The digitalization of Hard Copies Data will be entered properly into the field of Excel file which can be edited later if want and also it will create certain models on the basis of pattern it scans and later that analysis of model is done and the data is automatically processed to Excel format

REFERENCES

- [1] Nicolas Ragot Université François Rabelais Tours Laboratoire Informatique (LI EA6300) Tours, France nicolas.ragot@univ-tours.fr
- [2] Seong Jong Ha, Bora Jin, and Nam Ik Cho INMC, Department of EECS, Seoul National University, Seoul, S. Korea
- [3] Fanfeng Zeng, Guofeng Zhang and Jin Jiang, "Text Image with Complex Background Filtering Method Based on Harris Corner-point Detection"
- [4] Narasimha Murthy K N, Dr. Y S Kumaraswam, "Robust Model for Text Extraction from Complex Video Inputs Based on SUSAN Contour Detection and Fuzzy C Means Clustering", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011.
- [5] Rainer Herzog, Arved Solth and Bernd Neumann, "Text Block Recognition in Multi Oriented Handwritten Documents", Report, <http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/207/>, 2014.
- [6] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "ICDAR2013 Competition on Historical Book Recognition – HBR2013", 12th International Conference on Document Analysis and Recognition, 2013.
- [7] Zhixin Shi, Srirangaraj Setlur and Venu Govindaraju, "Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map", Proceedings of the 8th International Conference on Document Analysis and Recognition, pp. 794-798, Aug., 29- Sept. 1, 2005.