2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-9710-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAAIC53929.2022.9792982

A Geometric Method for Extracting Images of PDF Files

Ms. N Nagajothi
Ph D Research Scholar, Department of Computer Science
Alagappa University
Karaikudi, Tamil Nadu, India
nnjothi@gmail.com

Dr S S Dhenakaran
Professor, Department of C SC
Alagappa University
Karaikudi, Tamil Nadu, India
ssdarvind@yahoo.com

Ms S Uma Maheswari
Ph D Research Scholar, Department of C SC
Alagappa University
Karaikudi, Tamil Nadu, India
17umeshrani@gmail.com

***Abstract*— PDF is a format of the file that is used for showing research articles mostly. PDFs are actually easy to create, but extracting the data from PDF files is a challenging task. Images contain important as well as useful information which cannot be represented by text. The difficult concepts are better explained by an image than the text. In this paper, a feasible way to extract images from the PDF file is explored. This work is attempted to extract images from scholarly research publications of PDF files. The structure of a PDF file is entirely different from other file formats. A geometric method is proposed to get the position of an image in the PDF document and extracted all the images in that document. These images are saved in JPG or PNG or GIF or BMP format as a separate file. The resolution and size of extracted image are the same as the original image and are suitable for printing.**

Keywords— PDF, Image Extraction, Cartesian Coordinate System, User Space, Device Space

I. INTRODUCTION

PDF expansion is the Portable Document Format, which is used to display files in an electronic format in order to prevent unauthorized alterations to the file. It is software, hardware, and operating system agnostic. Text, photos and vector graphics, movies, animations, audio files, 3D models, interactive fields, hyperlinks, and buttons can all be included in the PDF format. These many components can be included in a PDF document to create an article, report, presentation, or portfolio.

PDF files are simple to make, read and use. However, PDFs are difficult to modify, and extracting data from them is a difficult operation. It is a way to show structured data. Research articles have wealth of information which is in PDF format only. Mostly process flow or algorithm or pseudo code or even formula can be represented in the form of images in research articles. It is not possible to extract texts or copy the images. So, a novel method is tried to explore images in the research articles that can be useful to researchers. It is seen that the retrieved images are good in size 96dpi as well as in good resolution. To include graphics in PDF, a device-independent Cartesian Coordinate-System is utilized to represent the surface of the page. Analytic geometry, sometimes known as Cartesian geometry, is a branch of Mathematics that describes every point in two-dimensional space using two coordinates, x, and y.

The x-axis is commonly defined or portrayed in Mathematics as horizontal and oriented to the right, while the y-axis is vertical and oriented upwards. The y-axis is directed downwards in Computer Graphics. Fig. 1. illustrates this. In a two-dimensional system, two coordinate spaces are maintained. The first is User Space, which is where graphics primitives are defined. The PDF page coordinate system is referred to as User Space. Device Space, on the other hand, is the coordinate system of an output device like a screen, window, or printer. Device space refers to the coordinate systemof where the images are drawn.

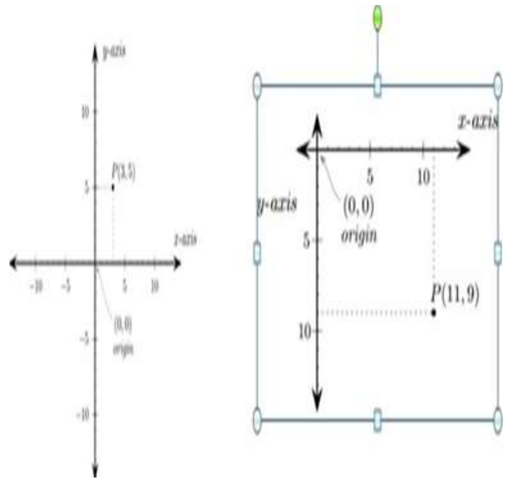


Fig. 1. Cartesian Coordinate System a.Mathematics Perspective b.Computer Perspective

The surface of the page is considered a two-dimensional space. Two coordinates are used for getting the location and size of the image on each page of the PDF files. The internal data structure of PDF is graphics state which is widely helpful in mapping user space to device space coordinate system.

The retrieved image size is larger than the raw image in the PDF file and stored in the systemfolder.

II. RELATED WORKS

A vast number of papers are now preserved in Portable Document Format (PDF) files, thanks to developments in information technology and extensive use of the Internet. For several years, document structure analysis has been studied (Mao et al., 2003) have provided a description of some approaches, and there are now a variety of freely available tools that can be used to extract information from scientific papers. Several information extraction methods had been introduced. "Scholarly figure" is a term used by Boschen et al., [1] to describe data visualizations such as plots or charts.

STALKER [2] is a method for extracting hierarchical information. The extracted data is separated into level, with lower levels containing more specific and concrete information than higher level.

CCWRAP [3] is another approach that separates data into tuples, with each tuple including certain attributes. LAPDFText [4] methodology only extracts text from research articles and is meant to be used as a starting point for more advanced extraction methods that handle multi-modal content such as images and graphs.

The system works in three stages: (a) by means of spatial layout dealing to set and recognize blocks of adjoining text, (b) by means of a rule-based system to organize text blocks into rhetorical categories, and (c) by edging categorized text blocks simultaneously in the exact arrange to take out text from section-wise clustered blocks. CERMINE [5] is a comprehensive system that can extract not only the metadata of a document, but also the content of the document, but also bibliographic references, and associated metadata, as well as the document's structured body content, directly from a PDF file.

Piotr Adam Praczyk et al. [6] have proposed the layout of a PDF page and understood a particular division of a page into the area called columns. Each area on a page is a sum of disjoint rectangles. This can be defined as a rectangle, let it be P, representing the page. The set D comprising subareas of a page in the document is called a page division.

$$\bigcup_{Q\in D} Q = P$$

$$\forall_{x,y\in D}\, x \cap y = \emptyset$$

$$\forall_{Q\in D}\, Q \neq \emptyset$$

$$\forall_{Q\in D}\, \exists_{R=\{x:x \text{ is a rectangle}, \forall_{y\in R\setminus\{x\}}\, y\cap x=\emptyset\}}\, Q = \bigcup_{x\in R} x$$

Graphical areas detected by a simple clustering do not directly resemble to images. The main reason for this is that images may contain not only graphics but also portions of text. For instance, common graphical elements not belonging to a figure include logos and text separators like lines and boxes and some parts of mathematical formulas include graphical operations. So this graphical layout should not be considered as an image.

Mingyan Shao et al.[9] have proposed a system that can extract and analyze figures from PDF documents and classify them using machine learning and built an interpreter that fostered to a sequence of self-contained graphic objects mirroring the PDF content stream. This system mainly focused on vector-based diagrams. But mostly images published in digital articles are raster-based. Document Image Analysis defined by George Nagy [10] is the concept of theory and practice of recovering the symbol structure of digital images scanned from paper or produced by a computer system.

Lawrence O'Gorman [11] described the Docstrum method for page layout analysis. This method can be used to segment independently oriented smaller documents like receipts, index cards, and business cards in a single image. Structural blocks are groups of one or more text lines assembled on the basis of spatial and geometric characteristics. These structural blocks ate determined by finding regions in contour line patterns [12],[13]with properties such as parallelness, perpendicular proximity, and overlap.

Kanungo and Haralick [14] proposed a method named geometric registration for aligning the OCR (Optical Character Recognition) tags on a page produced from a file of the text. This method produces the exact character locations and tolerates for degradation through copying and scanning the printout but contrasting string matching techniques, it applies only to synthetically generated pages.

A Cartesian coordinate system, as explained by Manish R. Joshi et al. [15], translates a pair of numbers to a particular position on a two-dimensional plane. A data position (u,v) is a pair off of points with a space "u" beside the x-axis and a detachment "v" corresponding to the y-axis revealed simultaneously. The x-axis and y-axis distances have symbols.

Jian Zhang et al. [16] described the method to consider the relative difference of means of pixel neighborhood to be the feature of every pixel due to ants' movement is decided by the relative difference of means of pixel circle neighborhood.

III. Technical Details of PDF

Text, vector graphics, and bitmap or raster graphics make up a PDF file. Vector graphics are illustrations and designs made up of shapes and lines. Photographs and other forms of images are represented using raster or bitmap graphics. To depict the surface of a page, PDF graphics use a device-independent Cartesian coordinate system.

A. Coordinate System of PDF

PDF contents are represented in the two- dimensional coordinate system. Each coordinate can be represented as a vector which is an ordered pair(x, y) or column matrix. The origin of the PDF coordinate system (0,0) denotes the bottom-left corner of the PDF page. PDF file specifies 72 points to 1 physical inch.

One important data structure is the graphics state that is associated with a PDF file. The Graphics state holds the information on how graphics are rendered to the screen. Its values are graphics control parameters such as the Current Transformation Matrix (**CTM**) and Color Space. The graphical parameter Current Transformation Matrix determines the coordinate system. These parameters are used to map from the user coordinate system into the device coordinate system when extracting the images from PDF pages.

IV. Methodology

Images are stored as a separate object XObject which contains the raw binary data for the image in the PDF file. However the image is part of an x-object container, the image location is not available. First, the PDF file is parsed to create a sequence of object instances. Next text stripped pages are created. Then it is a checked object which is an image object that was extracted with the help of a Cartesian coordinate system with two parameters x and y-axis values. Here it is considered user space to find the coordinates of images in the PDF pages and device space to fix the size of the extracted images according to the monitor or screen.

In order to find the exact size of the image, the x coordinate is increased to the right, and the y coordinate is increased downwards. The conversion between user space and device space is performed automatically during extraction. Because the PDF page coordinate system is user space. It is necessary to refer coordinate system of the screen or window which is device space. The targeted device is a screen or window. This was helpful in finding the location and size of the image object in the PDF file. Graphics state is the one important component of our work. It is associated with PDF files. The key factor of graphics state is the Current Transformation Matrix used to map from user space coordinate system that denotes PDF page coordinates to device space coordinate system which denotes targeted device which is monitor or screen.

A. Proposed Methodology

- Collect Research papers that are in PDF
- Apply parsing method to create a sequence of object instances
- Create text stripped pages
- Detect graphics contents in the text stripped pages
- If graphics contents are detected, extract images from those pages with the help of a two-dimensional Cartesian coordinate system
- Create an image file for each extracted image which may be JPG or GIF or PNG or BMP and store it in the specified computer folder path as a separate file

The above-proposed methodology is depicted in Fig. 2. The proposed system is to extract all the images in the entire PDF file and created separate image files that are stored in the folder directory. The extracted image size is somewhat larger than the raw image. Because device space is a device-dependent coordinate system. So extracted image size varies from the raw image which is on the PDF page.

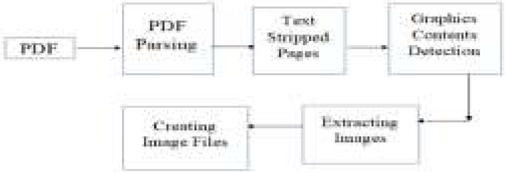


Fig. 2. Block Diagram Illustrating Proposed Approach

The resolution of the extracted image is 96dpi which has a good resolution for displaying the image on the monitor. The proposed method has extracted high-resolution images by doing the mapping process from user coordinates to devicecoordinates using CTM.

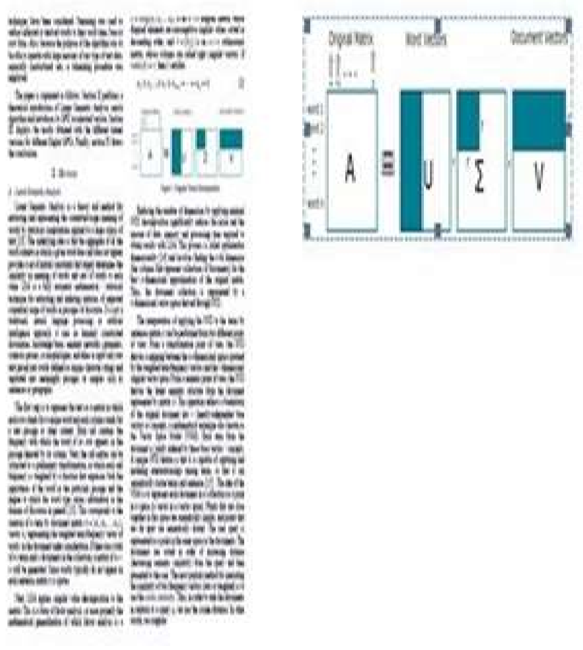


Fig. 3. Sample Page(Alexandru Iacob et al., [7]) and Extracted Coloured Image

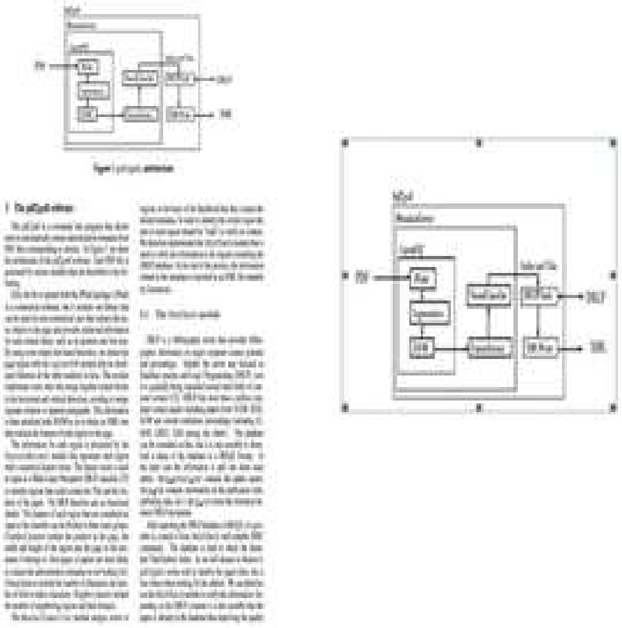
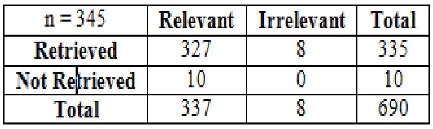


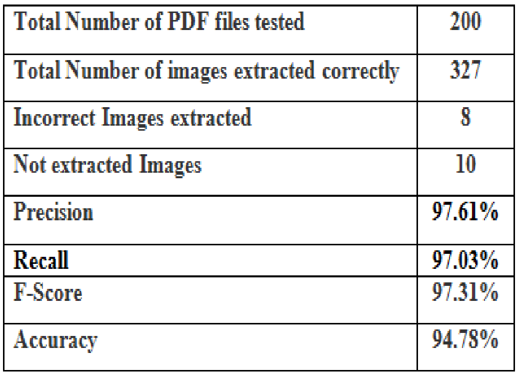
Fig. 4. Sample Page(Simone Marinai[8]) and Extracted Black and White Image

The proposed work can extract black and white as well as colored images from PDF pages. Extracted images are the same as original images which are in the PDF document. Fig. 3. and Fig. 4. show the PDF pages and extracted images. The proposed approach is with 200 PDF files. It has extracted 196 images from PDF files. This extraction procedure is complex due to the structural format of the PDF files. But images are more memorable than the text. It is interesting that most of the results are satisfactory from the user's point of view. These images can be used in a variety of ways.

V. RESULTS AND DISCUSSION



n = 345	Relevant	Irrelevant	Total
Retrieved	327	8	335
Not Retrieved	10	0	10
Total	337	8	690



Total Number of PDF files tested	200
Total Number of images extracted correctly	327
Incorrect Images extracted	8
Not extracted Images	10
Precision	97.61%
Recall	97.03%
F-Score	97.31%
Accuracy	94.78%

The performance of the proposed system on 200 PDF documents is focused on Computer Science files which are published in 2009-2014 years. These documents contain 345 images.

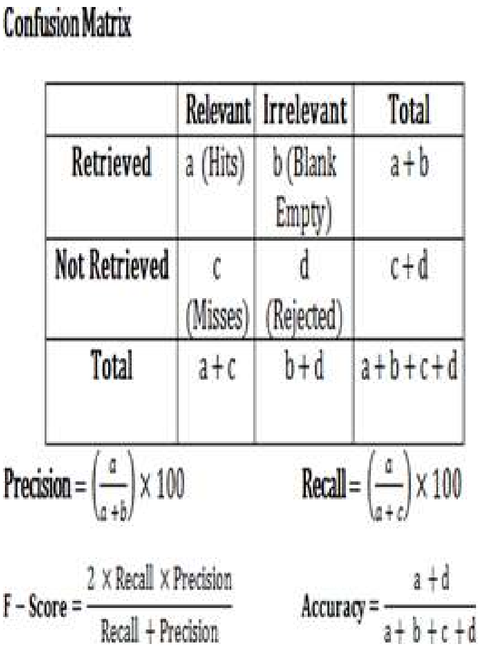
The performance of the new system is assessed by two methods such as one is execution (response) time and another one is standard performance metrics (Precision, Recall, F-Score, and Accuracy).

For the first method, the response time is calculated on two different machines. The proposed work is executed in two different machines. For one input, the First machine with 8GB RAM took a response time of 0.35seconds and the next mac For another input, the First machine with 8GB RAM took 42 seconds, and the next machine with 4 GB RAM took 44 seconds. The system time can differ by frequency of the clock, the frequency of the clock must be considered while calculating time. Factors affecting response time are,

a. Program b. Compiler c. Instruction Set Architecture d. CPU Design (Organization) e. Technology (VLSI) f. Data Set (PDF Document). g. system clock. So the response time varies from one input to another input as well as different configurations of machines.

In the second method, for evaluating the extraction applied factor is the standard metrics of precision, recall, F-Score, and Accuracy by using the confusion matrix which is defined in below Table I. It is seen that the performance is better than the existing method by producing an acceptable Accuracy value.

TABLE I. STANDARD METRICS OF PRECISION, RECALL,F-SCORE AND ACCURACY BY USING THE CONFUSION MATRIX



Confusion Matrix

	Relevant	Irrelevant	Total
Retrieved	a (Hits)	b (Blank Empty)	a+b
Not Retrieved	c (Misses)	d (Rejected)	c+d
Total	a+c	b+d	a+b+c+d

$$\text{Precision} = \left(\frac{a}{a+b}\right) \times 100 \qquad \text{Recall} = \left(\frac{a}{a+c}\right) \times 100$$

$$\text{F - Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad \text{Accuracy} = \frac{a+d}{a+b+c+d}$$

The results are tabulated in Table II. Which are obtained on executing the test data set. This experiment is conducted with nearly 200 PDF documents which have produced Precision 97.61%, Recall 97.03%, F-Score 97.31%, and Accuracy 94.78%. It is found that the proposed system has performed better in extracting images from the PDF documents producing quality images.

TABLE II. RESULTS OF TEST EXECUTION

Fig. 5. shows the graph of our system performance metrics.

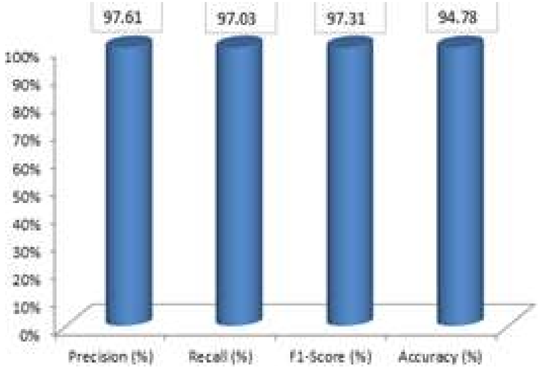


Fig. 5. PERFORMANCE OF IMAGE EXTRACTION

VI. CONCLUSION

An image or a picture is worth a thousand words. A single image can convey complex and multiple ideas. The images in the scientific articles are very much helpful to the researchers. Complex concepts are better explained by an image than the text. The PDF file format is an entirely different structure and uses a different coordinate system. So extracting text and images from PDF files is a complex task. This work is a foundation for extracting all images from the PDF files. The proposed method has parsed the PDF file and created the text stripped pages. This process is helpful to exactly locate the images. Here it is achieved high accuracy and also all the performance metrics are high. The resolution of the extracted image is 96dpi that is suitable resolution for displaying the image in the screen. A high resolution images are extracted by doing the mapping process from user coordinates to device coordinates using CTM. It is seen that this approach has taken less response time in high end machines. The future work is to use optimized machine learning algorithm for extracting desired images as well as achieve higher resolution rate which is suitable for printing.

REFERENCES

[1] *F. Böschen, T. "Beck, and A. Scherp, Survey and empirical comparison of different approaches for text extraction from scholarly figures", Multimedia Tools Appl., vol. 77, no. 22, pp. 29475–29505,2018*

[2] *Ion Muslea, Steve Minton, and Craig A. Knoblock. "A hierarchical approach to wrapper induction", Proceedings of the Third International Conference on Autonomous Agents, Seattle, pp. 221-227,1999*

[3] *Nicholas Kushmerick, "Wrapper Induction: Efficiency and expressiveness", Artificial Intelligence, pp.15-68, 2000*

[4] *Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, Gully APC Burns, "Layout- aware text extraction from full-text PDF scientific articles", Source Code for Biology and Medicine, Article Number 7, 2012*

[5] *Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M., and Bolikowski, L, "Cermine – Automatic Extraction of Metadata and References from Scientific Literature. In Document Analysis Systems (DAS)", 11th IAPR International Workshop, pp. 217–221. IEEE,2014*

[6] *Piotr Adam Praczyk, Javier Nogueras- Iso, and Salvatore Mele, "Automatic Extraction of Figures from Scientific Publications in High-Energy Physics", Information Technology And Libraries ,December 2013*

[7] *Alexandru Iacob, Lucian Itu, Lucian Sasu, Florin Moldoveanu, Cosmin Ioan Nita, Ulrich Foerster, Constantin Suciu," GPU Accelerated Semantic Search Using Latent Semantic Analysis",IEEE, 2016*

[8] *Simone Marinai, "Metadata Extraction from PDF Papers for Digital Library Ingest", 10th International Conference on Document Analysis and Recognition, pp. 252-255, IEEE, 2009*

[9] *Mingyan Shao and Robert P. Futrelle," Recognition and Classification of Figures in PDF Documents", Springer-Verlag Berlin, Heidelberg, 2005*

[10] *George Nagy, "Twenty Years of Document Image Analysis in PAMI, IEEE Transactions on Pattern Analysis and Machine Intelligence", Vol.22, No.1, January 2000*

[11] *Lawrence O'Gorman, "The Document Spectrum for Page Layout Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.15, No.11, November 1993*

[12] *L. O'Gorman and G. I. Weil, "An approach toward segmenting contour line regions", 8th Int. Conf. Patt. Recogn. (Paris), Oct. 1986*

[13] *M Seul, L R Monar, L O'Gorman, R Wolfe, "Morphology and Local Structure in Labyrinthine Stripe Domain Phases", Sci, Vol 254, Dec 1991*

[14] *T Kanungo and R M Haralick, "An automatic Closed-Loop Methodology for Generating Character Groundtruth for Scanned Documents", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.21, No.2, Feb 1999*

[15] *Manish R. Joshi, Yogita S. Patil," Analysis of change in coordinate system on clustering", IEEE,2016*

[16] *Jian Zhang, Kun He, Jiliu Zhou, Mei Gong, "Ant Colony Optimization and Statistical Estimation Approach to Image Edge Detection", IEEE,2010*

[17] *B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition", IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, 2017*

[18] *R.M. Haralick, "Document Image Understanding: Geometric and Logical Layout", Proc. Internet. Conf. On Computer Vision and Pattern Recognition, pp. 385-390,1994*

[19] *W.S. Lovegrove and D.F. Brailsford, "Document Analysis of PDF files: Methods, Results, and Implications", Electronic Publishing, Vol. 8(2 & 3), pp. 207-220,1995*

[20] *David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research 3, 2003*