IOP Conference Series: Materials Science and Engineering

Document Image Extraction System Design

Document Image Extraction System Design

N I Widiastuti[1*], K E Dewi[2]
[1,2]Informatics Engineering Department, Universitas Komputer Indonesia, Indonesia

Email : *Nelly.indriani@email.unikom.ac.id

Abstract. The design of the document image extraction system aims to provide an overview of the process in a system that converts the image of a document into text so that it is easier to use (save, manage or search for information) or representing in interesting visualization. The system design based on specific cases such as decision letters, certificates, and assignments. The document identified data that might be needed. After that, other processes are carried out based on each document. The design results show the flow of the system starts from data input, scanning, pre-processing to the document image, character classification, normalization process, and extraction process. This is done because the uniqueness of each document requires a unique process so that it cannot be treated in general. The design of the document image extraction system focuses more on the process of character recognition. The uniqueness of a document has an impact on the extraction process. Then the system design needs to be added to selected the type of document to be extracted.

1. Introduction

Almost every business or academic activity has a variety of documents used. Some of these documents are used to communicate such as incoming or outgoing letters and some documents are used as proof of activities such as certificates or decrees. An institution or individual, usually do the archiving of these documents. In general, archiving is done by scanning the document one by one. For some people, this is one way to apply the concept of digitalization. With this understanding, problems arise because documents cannot be traced or retrieve information. This is because the scanned file is unstructured. So we need a way to extract these documents into text so that information is easier to find when needed. In this case, it means changing the scanned file to a structured object such as a record in a database[1].

The concept of digitizing documents, in general, is an attempt to convert printed documents into digital form. The digital form in question is character type data so that it can be easily recovered (retrieve). This change process is not a simple thing because it involves character recognition technology or better known as Optical Character Recognition (OCR) and machine learning to classify[2]. After changing a digital document into a set of characters, each part of the document will be classified according to needy utilizing clean semantics from a structured database[3].

Some research has been carried out starting from extracting handwritten documents by Ranju Mandal et al. They try to recognize the date format written by hand. They use Dynamic Time Wrapping (DTW) and a feature-based approach to recognize month names instead. Also recognized by the gradient-based features and Support Vector Machines (SVM) such as numeric or other symbols.[4]. A survey was also conducted in 2000 by George Nagy in articles discussing Document Image Analysis (DIA) in an article published in the IEEE of 99 articles. The article discusses the types of documents that are the object of discussion, the cases resolved and the topics that support each research. the conclusion obtained is that the Hidden Markov Model (HMM) is the most widely used, the result document format is widely used XML. But in general, a good test is still a question. Appropriate input data for all cases have not yet reached an agreement to obtain comprehensive testing results [5]. In another article, Nagy et al. submit a system design that can analyze scientific article documents [6]. In research conducted by Kenny Wong et al. A prototype system can encode printed documents into digital [2].

Based on research that has been done, each discusses the methods needed to analyze the image of a document, but not specifically discussed the type of document that has a uniqueness each will require different handling. This research discusses the design of a system that extracts images of several types of documents that usually need to be extracted such as assignments, decrees, and certificates. Also, the system requirements for each document type will be discussed. This system design is useful for the process of development or implementation so that it can find back the information contained in the document image.

2. Method

This study discussed the system design that extracts a scanned document into data that can be easily recovered because it has been structured. The design of this system was formed based on superior research results in their respective problems. From the various studies the preprocessing method, character recognition, and segmentation methods were chosen, while good machine learning was also discussed for the classification of each part of the document. In the final section, the system architecture is depicted

3. Results and Discussion

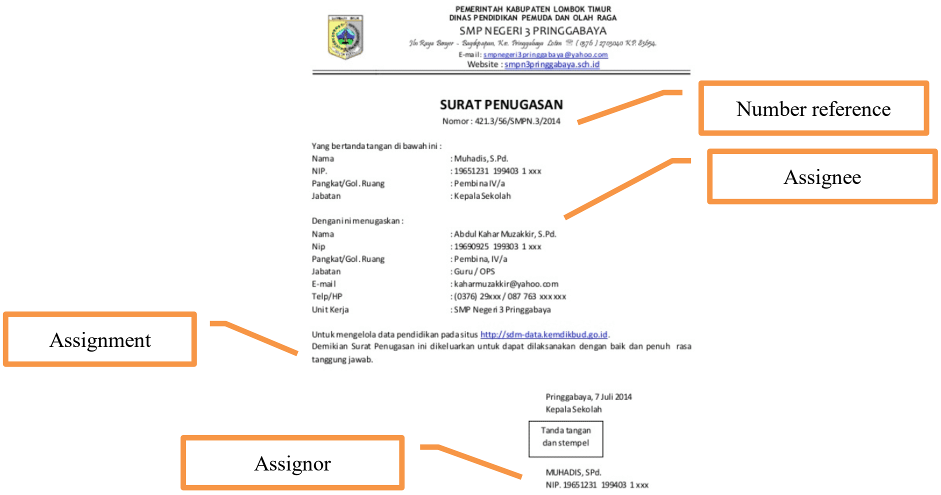
In this study Document Image Analysis (DIA) will be called document image extraction because it will be more emphasized on the purpose or final benefit of the DIA results, namely obtaining information from document images. The design of the document image extraction system will be divided into 4 stages. The first stage starts from processing the printed document into a file, the second stage then processes the image of the document or the pre-processing stage, the third stage is the process of character classification using machine learning and the fourth stage of the classification process of each part of the document according to user needs.

3.1. Documents

The documents in this study form the basis of the system design. This happens because the document is the data needed to find the information. In this study, the documents to be processed are documents that are usually uploaded as proof of lecturer performance. In general, these documents take the form of a letter of assignment, certificate, and decree. The document has parts that are considered important or commonly needed to be found again. This section will discuss the parts of the document.

3.1.1. Letter of Assignment.

This document is usually used as proof that every activity carried out is based on orders or tasks given by superiors. Information deemed important in the assignment letter is the letter number, who gave the assignment, to whom the assignment was given, what assignment was given, the assignment period and when the assignment letter was issued. In Figure 1. Important elements are shown in a letter of assignment



PEMERINTAH KABUPATEN LOMBOK TIMUR
DINAS PENDIDIKAN PEMUDA DAN OLAH RAGA
SMP NEGERI 3 PRINGGABAYA
Jln Raya Banyer - Bagekpapan, Kec. Pringgabaya Lotim ☎ (0376) 2703040 K.P. 83654.
E-mail: smpnegeri3pringgabaya@yahoo.com
Website : smpn3pringgabaya.sch.id

SURAT PENUGASAN
Nomor : 421.3/56/SMPN.3/2014

Yang bertandatangan di bawah ini :
Nama : Muhadis, S.Pd.
NIP. : 19651231 199403 1 xxx
Pangkat/Gol. Ruang : Pembina IV/a
Jabatan : Kepala Sekolah

Dengan ini menugaskan :
Nama : Abdul Kahar Muzakkir, S.Pd.
Nip : 19690925 199303 1 xxx
Pangkat/Gol. Ruang : Pembina, IV/a
Jabatan : Guru / OPS
E-mail : kaharmuzakkir@yahoo.com
Telp/HP : (0376) 29xxx / 087 763 xxx xxx
Unit Kerja : SMP Negeri 3 Pringgabaya

Untuk mengelola data pendidikan pada situs http://sdm-data.kemdikbud.go.id.
Demikian Surat Penugasan ini dikeluarkan untuk dapat dilaksanakan dengan baik dan penuh rasa tanggung jawab.

Pringgabaya, 7 Juli 2014
Kepala Sekolah

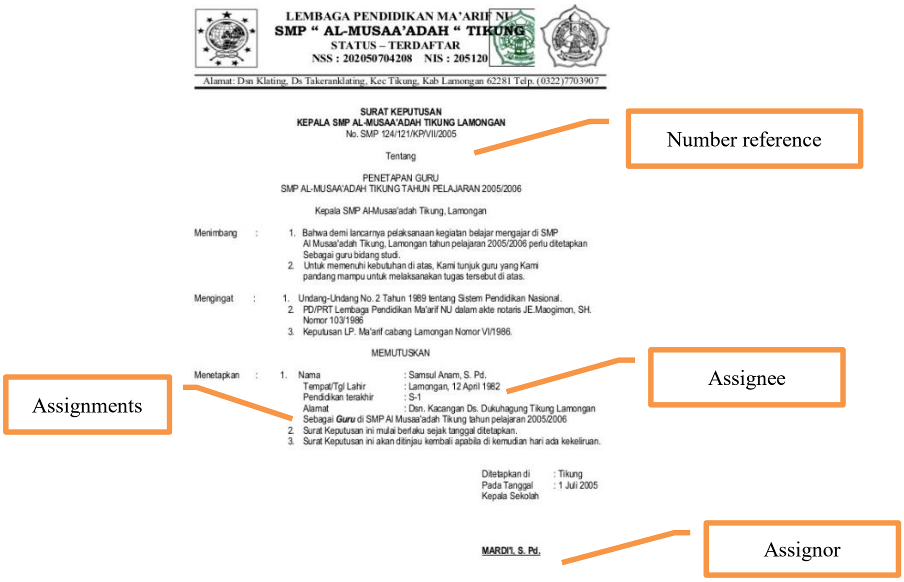
Tanda tangan
dan stempel

MUHADIS, SPd.
NIP. 19651231 199403 1 xxx

Figure 1. The Element of Letter of Assignment

3.1.2. Decree.

This document is usually given to someone who has changed functional, position or level. An example of a decree is a promotion letter. The usual element of information needed is the letter number, for whom the decision was made, who set it and when it was determined. Figure 2. shows the elements of information that are normally needed in the decision document.



LEMBAGA PENDIDIKAN MA'ARIF NU
SMP " AL-MUSAA'ADAH " TIKUNG
STATUS – TERDAFTAR
NSS : 202050704208 NIS : 205120
Alamat: Dsn Klating, Ds Takeranklating, Kec Tikung, Kab Lamongan 62281 Telp. (0322)7703907

SURAT KEPUTUSAN
KEPALA SMP AL-MUSAA'ADAH TIKUNG LAMONGAN
No. SMP 124/121/KP/VII/2005

Tentang

PENETAPAN GURU
SMP AL-MUSAA'ADAH TIKUNG TAHUN PELAJARAN 2005/2006

Kepala SMP Al-Musaa'adah Tikung, Lamongan

Menimbang :
1. Bahwa demi lancarnya pelaksanaan kegiatan belajar mengajar di SMP Al Musaa'adah Tikung, Lamongan tahun pelajaran 2005/2006 perlu ditetapkan Sebagai guru bidang studi.
2. Untuk memenuhi kebutuhan di atas, Kami tunjuk guru yang Kami pandang mampu untuk melaksanakan tugas tersebut di atas.

Mengingat :
1. Undang-Undang No. 2 Tahun 1989 tentang Sistem Pendidikan Nasional.
2. PD/PRT Lembaga Pendidikan Ma'arif NU dalam akte notaris JE.Maogimon, SH. Nomor 103/1986
3. Keputusan LP. Ma'arif cabang Lamongan Nomor VI/1986.

MEMUTUSKAN

Menetapkan :
1. Nama : Samsul Anam, S. Pd.
Tempat/Tgl Lahir : Lamongan, 12 April 1982
Pendidikan terakhir : S-1
Alamat : Dsn. Kacangan Ds. Dukuhagung Tikung Lamongan
Sebagai ***Guru*** di SMP Al Musaa'adah Tikung tahun pelajaran 2005/2006
2. Surat Keputusan ini mulai berlaku sejak tanggal ditetapkan.
3. Surat Keputusan ini akan ditinjau kembali apabila di kemudian hari ada kekeliruan.

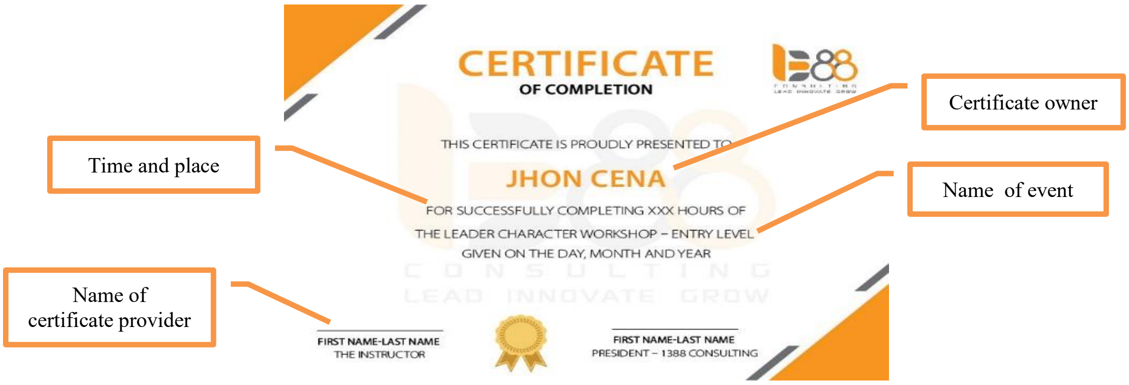
Ditetapkan di : Tikung
Pada Tanggal : 1 Juli 2005
Kepala Sekolah

MARDI'I, S. Pd.

Figure 2. Element of the Decree

3.1.3. Certificates.

In this case, the certificate in question is a document that proves the owner's participation in an activity. An important part of a certificate document is usually who owns the certificate, who gives it, the name of the activity or event, when it will be held, the certificate number (optional). Figure 3. shows the elements that are usually needed in a certificate.



CERTIFICATE
OF COMPLETION

THIS CERTIFICATE IS PROUDLY PRESENTED TO

JHON CENA

FOR SUCCESSFULLY COMPLETING XXX HOURS OF
THE LEADER CHARACTER WORKSHOP – ENTRY LEVEL
GIVEN ON THE DAY, MONTH AND YEAR

FIRST NAME-LAST NAME
THE INSTRUCTOR

FIRST NAME-LAST NAME
PRESIDENT – 1388 CONSULTING

Certificate owner
Time and place
Name of event
Name of
certificate provider

Figure 3. Element of the Certificate

3.2. *Pre Processing*

The essence of pre-processing is the stage that converts a printed document into a collection of characters that will be recognized in the OCR process. The process consists of grayscale, binarization, thinning, background removing, segmenting and scaling. These six processes are usually carried out to prepare the image for feature extraction so that it is ready to be classified. In this study, all documents require the process of greyscaling, binarization, segmenting and scaling. While background removing and thinning are only done on certificate documents.

Some research that has been done in binarization and segmenting will be reviewed as a recommendation of the method used in the design of this system. Caballero et al's research results can be used in the design of this system to set the threshold in the binarization process. The threshold is learned through selecting the lightness or hue component of the color input cell, increasing the quality of the bitmap, and calculating the segmentation threshold range for this cell. Then, starting from the study threshold range in each display cell, the best threshold for each cell is obtained[7]. Other studies conducted by A. Cheung et al. use segmentation to recognize characters in Arabic[8]. The segmentation method they used to separates the overlapping Arabic characters horizontally then is repeated to control the combination of segments. This method is required for the certificate documents. In the next segment, specific pre-processing is explained for each document.

3.2.1. *Letter of Assignment.*

In processing this document it has its difficulty level. The thining process was not performed in this document because the font size is very small. Thining will result in the loss of the character. Likewise, scaling, this process is not required for font sizes that are already small. In the segmentation process, letters of assignment documents require a good method for cases of small font sizes or have the possibility of documents being damaged due to age or an imperfect scan process. Segmentation is essentially cutting the document into small pieces that will be considered as characters.

This study suggests that the segment method used by Qarouse et al. The segmenting process is done using two-level segmentation. The first level of segmentation is done to get the word. The method used is profile projection and Interquartile Range (IQR) to separate words from characters in words. In the second level, projection profiles are used together with statistical and topological features. The results show a satisfying level of accuracy[9].

3.2.2. *Decree.*

The pre-processing for this document is almost the same as the letter of assignment. The main problem in this document is the format of the document that contains the table. In several decrees given to several people at once. This results in the detection of decision recipients requiring special pre-processing.

A study conducted by Yuri Tijerino et al. Generates a table mapped to ontology using text-based features [10]. In the following year by David W. Embley et al conducted a survey of research that discusses table extraction. [11]. They conclude that research that has been done is still focused on overcoming the grid rather than extracting the contents of each cell. Research conducted by Ramel et al. proposes several ways to extract tables based on their design[12]. In this study, documents that have table elements will not be addressed because they require special discussion.

3.2.3. Certificates.

This document has special characters because many have picture elements. Besides the type of font used in the certificate is very diverse both in terms of size or type, so that requires a variety of training data as well. Specifically, the thining process is only done for certificate document images because of the large character size.

In certificate document images, backgrounds or logos often appear and cause character detections to experience errors. For background problems, we found several studies to overcome this. In a study conducted by Liwen Gao et al. eliminate background with the OTSU segmentation method[13]. Wei Fang et al. remove the background by modifying the convolution neural network (CNN))[14].

3.3. Optical Character Recognition

This stage is the process of classifying images into character classes that are recognized by machine learning. Almost all supervised machine learning can classify characters. However, attention needs to be paid to the preprocessing that can reduce the problems that may arise.

Several studies have suggested some algorithms including those conducted by Phangtriastu et al who compared artificial neural networks (ANN) with support vector machines (SVM). Their research shows that the SVM and zoning and profile projection extraction features have a high degree of accuracy[15]. For certificate document images, special handling of cursive fonts such as in research conducted by Saeeda Naz et al. Their research emphasizes the introduction of segmentation[16].

3.4. Document Image Extraction

The complete document image extraction system design can be seen in Figure 4. In the design, the documents to be processed are specifically explained. This is also the system limit that will be applied. Each document has several differences in pre-processing, especially certificates. The certificates specifically experience background removing, segmentation and thining. While the other two documents, after gray scaling and binarization, then segmentation. Scaling is done to uniform the size of each character.

After the extraction and classification features, the OCR process has finished. Then the information extraction process is needed based on each document. In this process, the data obtained is already in the form of characters. This data goes through a process of normalization that aims to ensure each character is reorganized into words that have meaning. Ways that can be used in normalization such as string matching or word similarity. Where the core of this process is comparing data tokens with a word dictionary. Some classes such as certificate owners or assignees do not need to go through this process because it will not be in the dictionary.

To complete the extraction process, it is necessary to determine the features that are suitable for each class. In this study, we determine the features used are words. For example for certificate documents, the certificate owner class will appear after the word "to". Another example of decree documents, the class of decision recipients will appear after the words "decide" and "Name:" as well as for other classes. In the case of documents used, these features are very specific. For this reason, we use knowledge-based classifications.

In this study, we chose to study text from character images or OCR. Correcting text is relatively easier and the level of computing is light. In contrast to the following studies. In the system built by K. Y Wong et al. The system is created using the run-length method to arrange images from text. They used two evaluations, namely discussing segmentation to replace the text with images while asking the two to conduct training on various types and sizes of characters[2]. A similar method is used by Chengpei Xu et al. The difference lies in the object of research consists of video lectures to be generated into presentation slides [17]. Also, the results of a paper survey conducted by Nagy et al. 99 scientific articles that discuss the image of OCR problem analysis documents are still a challenge[5]. This second article finds findings on the problem of segmentation in text images.

The main purpose of making this system design is to obtain important information in each document that can be manipulated. The features used in the design of this document's image extraction system are the general features that appear in each document. The problem with OCR results is not good, we overcome by looking for words that have proximity to words in the dictionary. In previous studies such as those conducted by Nagy [5] or K. Wong[2] more focused on the conversion of document images into text. But in the research of George Nagy extracting in scientific journals by finding the location of each part to be stored information[6].

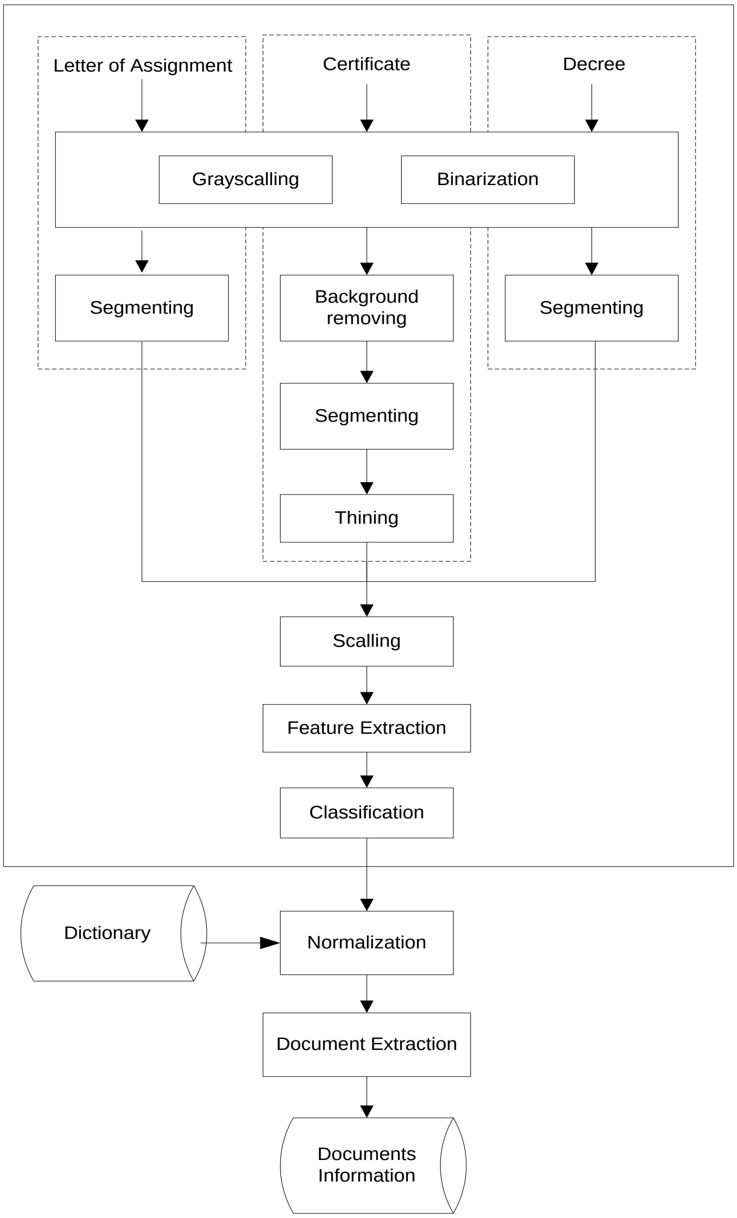


Figure 4. Documents Image Extraction System Design

4. Conclusion

The design of the document image extraction system has been described in this article. This design can describe what processes are needed to extract the image of a documentary letter of assignment, certificate, and decree. There are two core systems, namely the OCR section and information extraction. This design can be used to implement a document extraction system for a person's performance in a job. But this design still needs to be tested for success in detecting characters from different document images and the suitability of word features for knowledge-based extraction.

Acknowledgment
Thank you to UNIKOM for funding this research in 2019 funding.

References

[1] H. Saggion, T. Poibeau, J. Piskorski, Yangar, and R. Ber, 2013. *Multi-source, Multilingual Information Extraction and Summarization*. Berlin: Springer-Verlag, 2013.

[2] Wong, K. Y., Casey, R. G., & Wahl, F. M. 1982. Document analysis system. *IBM journal of research and development*, ***26***(6), pp.647-656.

[3] Sarawagi, S. 2008. Information extraction. *Foundations and Trends® in Databases*, ***1***(3), pp.261-377.

[4] Mandal, R., Roy, P. P., Pal, U., & Blumenstein, M. 2015. Multi-lingual date field extraction for automatic document retrieval by machine. *Information Sciences*, ***314***, pp.277-292.

[5] Nagy, G. 2000. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (**1**), pp.38-62.

[6] Horst, B. 2003. Recognition of cursive Roman handwriting-past, peresent and future. In *Proceeding of the seventeen international conference on document analysis and recognition ICDAR, In* (pp. 448-459).

[7] Fernández-Caballero, A., López, M. T., & Castillo, J. C. 2012. Display text segmentation after learning best-fitted OCR binarization parameters. *Expert Systems with Applications*, ***39***(4), pp.4032-4043.

[8] Cheung, A., Bennamoun, M., & Bergmann, N. W. 2001. An Arabic optical character recognition system using recognition-based segmentation. *Pattern recognition*, ***34***(2), pp.215-233.

[9] Qaroush, A., Jaber, B., Mohammad, K., Washaha, M., Maali, E., & Nayef, N. 2019. An efficient, font independent word and character segmentation algorithm for printed Arabic text. *Journal of King Saud University-Computer and Information Sciences*.

[10] Tijerino, Y. A., Embley, D. W., Lonsdale, D. W., Ding, Y., & Nagy, G. 2005. Towards ontology generation from tables. *World Wide Web*, ***8***(3), pp.261-285.

[11] Embley, D. W., Hurst, M., Lopresti, D., & Nagy, G. 2006. Table-processing paradigms: a research survey. *International Journal of Document Analysis and Recognition (IJDAR)*, ***8***(2-3), pp.66-86.

[12] Ramel, J. Y., Crucianu, M., Vincent, N., & Faure, C. 2003. Detection, extraction and representation of tables. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* (pp. 374-378).

[13] Gao, L., & Lin, X. 2019. Fully automatic segmentation method for medicinal plant leaf images in complex background. *Computers and Electronics in Agriculture*, ***164***, pp.104924.

[14] Fang, W., Ding, Y., Zhang, F., & Sheng, V. S. 2019. DOG: A new background removal for object recognition from images. *Neurocomputing*, ***361***, pp.85-91.

[15] Phangtriastu, M. R., Harefa, J., & Tanoto, D. F. 2017. Comparison between neural network and support vector machine in optical character recognition. *Procedia computer science*, ***116***, pp.351-357.

[16] Naz, S., Hayat, K., Razzak, M. I., Anwar, M. W., Madani, S. A., & Khan, S. U. 2014. The optical character recognition of Urdu-like cursive scripts. *Pattern Recognition*, ***47***(3), pp.1229-1248.

[17] Xu, C., Wang, R., Lin, S., Luo, X., Zhao, B., Shao, L., & Hu, M. 2019. Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 898-903.