

Advanced content-based retrieval for digital correspondence documents with ontology classification

Rifiana Arief, Suryarini Widodo, Ary Bima Kurniawan, Hustinawaty, Faisal Arkan
Faculty of Computer Science & Information Technology, Gunadarma University, Depok, Indonesia

Article Info

Article history:

Received Nov 16, 2021
Revised Apr 19, 2022
Accepted May 21, 2022

Keywords:

Classification
Content
Document
Ontology
Retrieval

ABSTRACT

The growth of digital correspondence documents with various types, different naming rules, and no sufficient search system complicates the search process with certain content, especially if there are unclassified documents, the search becomes inaccurate and takes a long time. This research proposed archiving method with automatic hierarchical classification and the content-based search method which displays ontology classification information as the solution to the content-based search problems. The method consists of preprocessing (creation of automatic hierarchical classification model using a combination of convolutional neural network (CNN) and regular expression method), archiving (document archiving with automatic classification), and retrieval (content-based search by displaying ontology relationships from the document classification). The archiving of 100 documents using the automatic hierarchical classification was found to be 79% accurate as indicated by the 99% accuracy for CNN and 80% for Regex. Moreover, the search results for classified content-based documents through the display of ontology relationships were discovered to be 100% accurate. This research succeeded in improving the quality of search results for digital correspondence documents as indicated by its higher specificity, accuracy, and speed compared to conventional methods based on file names, annotations, and unclassified content.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rifiana Arief
Faculty of Computer Science & Information Technology, Gunadarma University
Margonda Raya 100, Pondok Cina, Depok, West Java, Indonesia
Email: rifiana@staff.gunadarma.ac.id

1. INTRODUCTION

The era of digital transformation and the availability of good-quality scanners and cameras currently support the recent trend of digitizing documents in different fields. The rapid growth associated with the use of digital documents means there is a need for a good archive system to ensure the documents are found easily when needed. This can be achieved by providing document names and annotation of the information on the document category and the attachment of relevant files. The documents are manually annotated using keywords and searched using file names or annotations/keywords [1], but this manual approach is highly ineffective and requires a longer time considering the continuous increase of digital documents. Therefore, automatic annotation is required to ensure efficient retrieval, organization, classification, and auto-illustration of digital documents [2]. This is necessary because of the problems associated with manual archiving such as the need to search for a document containing a certain text one after the other based on the name in the database [3]. Meanwhile, a continuous increase in the number of documents in the repository makes it unfeasible to use this method considering the difficulty associated with visualizing the content information from the documents subjectively. Inappropriate document annotation also leads to the discovery of irrelevant

search results [4]. It is important to reiterate that the increasing application of different types of digital correspondence documents with different naming rules and without a sufficient search system complicates the search process in certain content, specifically concerning unclassified documents, which makes the search process inaccurate and very slow [5]. It is possible to search documents through adopted content which involves matching the text-similarity instead of using the name or annotation to extract the textual content [6].

Document image retrieval is normally used to find the appropriate document image from the database based on the user's queries. It has two approaches which include retrieval based on text recognition using optical character recognition (OCR) which relies on recognizing the text from the document and later examining the similarity as well as the retrieval without text recognition (not using OCR) which relies on image features in the document and later calculating the similarity with the actual content of the images [7]. This simply means the content-based document image retrieval method has 2 main stages which are the extraction of the text/text image feature and the search based on matching the text/text image [8]. It is important to note that the search for the scanned documents based on the text similarity requires OCR with good accuracy to retrieve the results containing certain text according to the queries and ensure accurate matching and avoid errors. Tesseract OCR is currently the open-source tool with the best accuracy and availability in different languages which is normally used to recognize and extract texts from different image documents [9].

Several government agencies in Indonesia have digitized their documents such as the digital correspondence documents each of which generally has a document hierarchy that includes the classifications of the manuscripts of letters, types of letters, origins of letters, and subjects of certain letters according to the correspondence applicable to the institution. This digital correspondence document is usually archived through manual naming of the documents and the creation of annotations to classify the criteria which are further used in the search process. The system operator is expected to read the content and group or classify the document based on the type, origin, and subject while naming the document file to make its information more understandable. However, the increasing number of digital correspondence documents in the repository makes the use of this manual method of archiving and searching to be inefficient. It was also discovered that the difficulty in finding the required documents is due to the focus of the search process on file names and annotations which frequently produces non-required documents. Further exploration of the document's content is required to determine the information on the type of classification and hierarchy of the documents found and this usually makes the search process longer. This simply means that the government is faced with the problem of developing an appropriate archiving system of digital correspondence documents that can make it easy for the operator to classify documents automatically according to the document hierarchy and create an adequate annotation for classification while saving documents into the database. Another problem is the absence of a content-based search system according to the classification which can display the ontology relationship information and the document hierarchy. These are necessary considering the increasing number as well as the different types of digital correspondence documents currently being used in different organizations which require appropriate automatic archiving and retrieval systems to ensure the documents needed are found easily and accurately.

Several research [10]-[15] have been conducted about content-based image document search using OCR technology but they are only limited to searching base content for scanned documents without focusing on classified documents for a more specific search. Meanwhile, the increasing number and diversity of documents are making the classification process important to direct, summarize, and organize the documents easily, with efficient and cost-effective solutions [16]. Document classification is defined as the automatic grouping of documents into certain criteria based on similar content such as the subject, topic, language, field, and several others [17]. The rapid increase in the number of documents available in organizations has made it necessary to develop and implement an automatic document classification system to ensure effective management of text documents and large volumes of unstructured data as well as to retrieve information fast and accurately [18]. Some research has been conducted on this concept such as the classification of documents using different methods including support vector machine (SVM), k-nearest neighbors (KNN), Naive Bayes, and others [19]. Moreover, the continuous development and use of different forms of data, the validity of uncertain data forms, and the need for fast access have also led to the implementation of deep learning neural networks for classification due to their ability to exceed conventional machine learning methods in characterizing big data [20]. The use of this method is necessary because the performance of conventional methods usually reduces as the number of documents increases [21]. Convolutional neural networks (CNN's) deep learning method gives higher classification accuracy results than the Gauss Naive Bayes, random forest, Naive Bayes, and SVM methods [22] and example of this is the CNN applied by [23]-[25] which was observed to have accurately classified different documents with unstructured text content but was unable to classify those with special patterns such as specific and short strings. It also

requires many data and a longer time for the training and testing processes. Meanwhile, another research by [26]-[28] classified documents efficiently based on these special patterns and codes using the regular expression method which was quicker compared to the machine learning or deep learning method because it does not process training and testing data in advance. However, it cannot be used to classify different types of documents with unstructured text content. Another research by [29] automatically applied a multilevel classification method to scanned documents in image format based on the document hierarchy. OCR technology was utilized to recognize the text in these scanned documents in the form of digital correspondence documents and also to automatically classify the letters based on the document hierarchy up to a depth of 4 levels which include 4 classifications of the manuscripts of letters -> 5 classifications of the types of letters -> 15 classifications of the origins of letters -> 25 classifications of the subjects of letters by combining the CNN and regular expression method. This means the proposed method was able to classify documents with unstructured text into different criteria based on the types of the letters and also to categorize the contents of the letters in the form of special codes to origins and subjects of different letters based on special patterns including specific and short strings. However, the CNN method was only able to classify 5 types of letters with 94% accuracy while the Regular Expression pattern developed classified the origin, subject, and manuscript of letters with 100% accuracy. It was discovered that the method is only limited to new types of decision letters, assignment letters, invitation letters, and certificates which means it does have the capacity to classify all types of letters. It did not also form automatic ontological relations from the classification results and failed to conduct searches based on classified content by displaying the ontological relations despite the importance of the concept in resolving data heterogeneity through the provision of information on the document relationships and indicating strong support its retrieval based on the relevant content [30]. The research by [31] constructed an ontology from the classification of large-scale text data using the CNN to describe hierarchical information and document relationships but was not utilized to support the search for scanned documents, specifically digital correspondence documents, based on classified contents.

The overall review of previous research showed that content-based search has been widely applied to scanned documents using OCR but there is no specific focus on the classification as well as ontology construction as [32]-[34] to describe the hierarchical information on the scanned documents classified. Therefore, this present research is a significant improvement on [29] by developing classification capabilities for 5 to 22 types of documents. It also proposed a method to search classified content-based digital correspondence documents with automatic ontology relations to overcome problems associated with their searching process and search classified content by displaying ontological relationships to make it possible to easily trace the documents containing certain content and attribute information on the official manuscript, types, origin, and subject of the letters classified.

This advanced content-based retrieval system for digital correspondence documents with ontology classification (OCR-assisted) proposed two contributions which include document archiving with automatic hierarchical classification and content-based searching with more specific filters according to certain classification criteria as well as the provision of information on ontology relationships which shows the hierarchy including the manuscripts of letters -> the types of letters -> the origins of letters and the subjects of letters of each document found. This method is urgently required as a solution to the problems observed with conventional search systems of digital correspondence documents which are designed based on document names and annotations as well as unclassified content-based search.

2. METHOD

The object of this research is the digital correspondence document (which is non-confidential) belonging to the General Bureau of the Secretariat General of the Ministry of Education and Culture of the Republic of Indonesia. In total there are 22 types of letter, namely regulation letter, circular letter, procedure letter, decree, letter of instruction, warrant, letter of assignment, official memo, memo, invitation, cover letter, memorandum of understanding, letter of agreement, power of attorney, news, letter of statement, official letter, statement letter, announcement letter, report letter, notes and application letters (*in Indonesian Language: surat peraturan, surat edaran, surat prosedur, surat keputusan, surat instruksi, surat perintah, surat tugas, nota dinas, surat memo, surat undangan, surat pengantar, nota kesepahaman, surat kerjasama, surat kuasa, berita acara, surat keterangan, surat dinas, surat pernyataan, pengumuman, laporan, notula, dan surat permohonan*). There are 5 classifications of manuscripts of letters, namely manuscript of special service, manuscript of correspondence service, manuscript of determination service, manuscript of assignment service, and manuscripts of the regulatory service (*in Indonesian Language: naskah khusus, naskah korespondensi, naskah keputusan, naskah penugasan dan naskah peraturan*). Each letter has a letter-number in the form of a short string and a certain/specific pattern that can show information on the origins of letters and the subjects of letters. The classification for the origin of the letter and the subject of the letter is as

same as in the previous studies [29]. The advanced content-based retrieval system of digital correspondence document with ontology classification as indicated in Figure 1 consists of 3 stages: preprocessing, archiving, and retrieval.

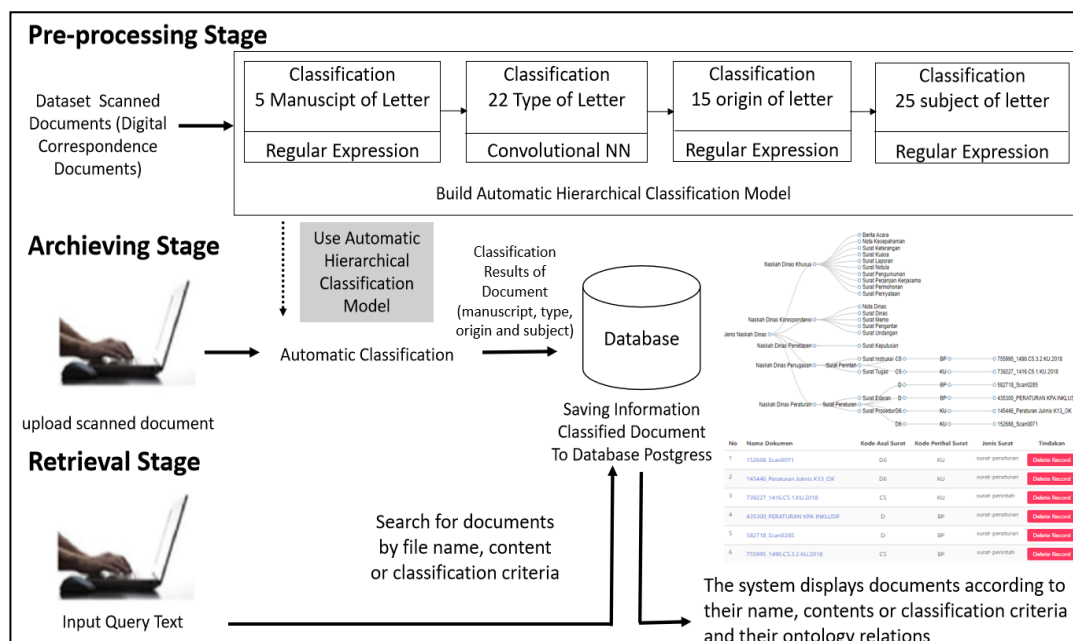


Figure 1. Advanced content-based retrieval system for digital correspondence documents with ontology classification

2.1. Pre-processing stage

The pre-processing stage involves the development of the automatic hierarchical classification model of digital correspondence documents based on the text classification approach in line with [29] that have been successfully classified digital correspondence documents into 4 manuscripts of letter, 5 type of letter, 15 origins of letter, and 25 subjects of letter with 94% accuracy by combining CNN and regular expression methods. The automatic hierarchical classification model developed in this research covers all the criteria according to the hierarchy of digital correspondence documents which include 5 manuscripts of letters, 22 types of letters, 15 origins of letters, and 25 subjects of letters. The manuscript, origin, and subject were classified using the same regular expression pattern applied in the previous research [29] while the types of letters were based on the CNN. Moreover, this model was further applied in the archiving stage. It is important to note that each letter (document) has a number in the form of a short string and a specific pattern showing information on its origin and subject letter.

The CNN architecture formed was used to classify the documents based on the type of letter and the process was initiated by configuring a basic neural network with epochs 20 followed by the configuration of the layer on the neural network by adding an input layer, convolution layer cnn3 with kernel size (3, 300), convolution layer cnn4 with kernel size (4, 300), and convolution layer cnn5 with kernel size (5, 300) with input from the input and output to feature layers, pooling layer with maximum type, dropout (0.5), combined layer (cnn3, cnn4, cnn5), output layer with loss function. MAXENT function, SOFTMAX activation function, input (3 * cnnLayerFeatureMaps), and 22 output. It is important to note that there are 22 classifications based on the types of letters while the manuscripts have 5 classifications which were used as the hierarchical tree of the types of letters in line with the modified Regex pattern adapted from [29].

The algorithm used in developing the automatic hierarchical classification model is as follows:

Input: Dataset Scanned Documents (Digital Correspondence Documents)

Output: Automatic Hierarchical Classification Model

(22 Types of Letters, 5 Manuscripts, 15 Origins, and 25 Subjects of Letter)

1. Prepare the training data and test data for 22 classifications of the types of letters.
2. Extract Scanned Documents to Text Using Tesseract OCR.
3. Create word vectors by taking all the word results from document extraction derived from the training data, and the test data from all types of classes.

4. Create Convolutional Neural Network architecture to 22 outputs of classifications types of letters.
5. Load a new word vector formed from the training data, the test data from 22 criteria of types of letters for the process of training and testing.
6. Train and test to get the best classification model for the types of letters (highest accuracy).
7. Use Regex patterns to classify the Origin and Subject of Letter adapted from [29].
8. Modify the Regex pattern to classify the Manuscripts of Letter adapted from [29]. Add the Regex pattern "Manuscripts of the Regulatory Service" and adjust the 5 classifications of the Manuscript of Letter as the hierarchical tree of 22 criteria of the types of letters.
9. Run an Automatic hierarchical classification model using a combination of Convolutional Neural Network + Regular Expression to classify the scanned document (Type of letter, Manuscript of letter, Origin of letter, and Subject of letter).
10. Save the automatic hierarchical classification model for use when archiving documents.

The algorithm to modify the classification algorithm for the manuscript of the letter is as follows:

Input: Classification Result based on Type of Letter

Output: Classification for Manuscript of Letter

1. Retrieve documents that have been classified according to Type of Letter Through the Convolutional Neural Network method.
2. Decide the pattern for the Manuscript Letter based on the Type of Letter.
 - If Type of Letter = Decree
 - Then Manuscript Letter = Manuscript of Determination Service
 - If Type of Letter = Letter of Instruction, Warrant, Letter of Assignment
 - Then Manuscript Letter = Manuscript of Assignment Service
 - If Type of Letter = Official Note, Official Letter, Memo, Invitation, Cover Letter
 - Then Manuscript Letter = Manuscript of Correspondence Service
 - If Type of Letter = News, memorandum of understanding, letter of statement, power of attorney, report letter, notes, announcement letter, letter of agreement, application letters, statement letter
 - Then Manuscript Letter = Manuscript of Special Service
 - If Type of Letter = Circular Letter, Regulation letter, Procedure Letter
 - Then Manuscript Letter = Manuscripts of the Regulatory Service
 - In addition, Manuscript Letter = No Category
3. Match the criteria for the Manuscript of Letter that is suitable based on the type of Letter.
4. Receive the suitable criteria for the Manuscript Letter.
5. Save the classification result for the Manuscript Letter from documents.

2.2. Archiving stage

The archiving stage was used to archive the classified documents into the database through two important stages. The first stage was the automatic classification of each scanned document file which is digital correspondence document according to their hierarchy using the automatic hierarchical classification model developed in the preprocessing stage. The second stage was to save the documents that have passed the automatic classification process to the PostgreSQL database.

The algorithm used to archive the documents through automatic classification is as follows:

Input: Scanned Document (Digital Correspondence Document)

Output: Classified Digital Correspondence Document stored in the PostgreSQL database

1. Upload the scanned document file to be archived.
2. Extract scanned documents to get text using Tesseract OCR.
3. Run the classification using an automatic hierarchical classification model that has been made (in Pre-processing stage) to obtain criteria according to the letter hierarchy.
4. Read the extraction results (document id, document name, document content) and the classification results (manuscript letter, type of letter, origin of letter, subject of letter).
5. Save the information (document id, document name, document content, classification results (manuscript letter, type of letter, origin of letter, subject of letter) into the table in PostgreSQL Database.

2.3. Retrieval stage

The search stage was used to find the digital correspondence document having the required text content and more specifically according to the required classification criteria including the type, origin, and subject of the letters as well as to display the information of ontology relationships formed based on the classification to ease the search process. This means the retrieval stage involves searching digital correspondence documents based on the classified content, forming the automatic ontology relationships from the classification results from the required letter, and displaying the appropriate digital correspondence document files and information of ontology relationships on the user interface based on the letter

classification. The search process started by entering a query in the form of text and selecting the classification depending on the content or letter classification criteria which is usually based on the name of the document by default. The text query was received by the controller for the search record process after which the keywords are sent to the data access object and the select command was run on the PostgreSQL database to search and filter content according to these keywords and required classification. The query results were in the form of an array list and the structure of the classified document relationships was automatically formed in the form of the classified document relation graph or chart on the search page. Meanwhile, the establishment of document relationships requires library data driver document and document object model that support large-scale data visualization to ensure quick and accurate description. Moreover, the query retrieved and counted all records in the database using an additional form of like to filter using the previously determined keyword parameters. The data filtered from the PostgreSQL database was used to establish a relationship between the document and the library through the data driver document and the digital correspondence documents after which the ontology relationships of the classification were provided. The appropriate digital correspondence document file was eventually displayed.

The Algorithm of advanced content-based retrieval system with ontology classification is:

Input: text on questions and options (document name (default), content or classification of letter)

Output: the document file found and the information of ontology relationships from letter classification

1. Receive requests in the form of text questions or keywords to search for digital correspondence documents by the controller in the processor from the input receiving unit in the user interface.
2. Search for records by the controller on the processor by the document name (default), content, or more specifically by letter classification.
3. Send a request by the controller in the processor to the data access object in the PostgreSQL database.
4. Execute the select command by the processor in the PostgreSQL database based on the query text and based on a search by to search and filter the content according to the query and the searched classification, so it results in an array list.
5. Form the relationship structure of the automatic classified documents by the processor according to the results of the request into the form of an ontology graph/relationship chart of the classified documents according to the results of the request to be displayed in the user interface.
6. Take and count all records in the database table based on a query by using an additional form of like to filter with predefined keyword parameters.
7. Form a document relation by the processor with a library provided by the data driver document.
8. Display the digital correspondence document found with the ontology relationships of such letter classification.
9. Select and display the required digital correspondence document.

3. RESULTS AND DISCUSSION

A total of 11000 digital correspondent documents in the form of scanned documents in PDFImages format were used and these include 8800 training data with each type of letter totaling 400 labeled documents and 2200 test data with each type of letter totaling 100 labeled documents used in developing the automatic hierarchical classification model. Moreover, unlabeled 100 new scanned documents other than those used for the training and test were used to experiment archiving process and the content-based search through the display of the ontology relationships.

The automatic hierarchical classification model developed was successfully used to classify a digital correspondence document automatically based on the hierarchy of 5 manuscripts of letters \Rightarrow 22 types of letters \Rightarrow 15 origins of letters \Rightarrow 25 subjects of letters. The result for the classification of the 22 types of letters using the CNN method was evaluated for 20 epochs with the accuracy, precision, recall, and F1 Score presented in Table 1.

Table 1. Evaluation result for the 22 types of letters CNN

No	Epoch	Accuracy	Precision	Recall	F1Score
1	Epoch 1	0.4091	0.4213	0.4084	0.3802
2	Epoch 5	0.6517	0.6834	0.6516	0.6425
3	Epoch 10	0.7005	0.7269	0.7005	0.6918
4	Epoch 15	0.7340	0.7611	0.7339	0.7237
5	Epoch 20	0.7540	0.7765	0.7540	0.7415

Table 1 shows the inaccurate results recorded for these types of letters are possibly associate with (1) the limitations in obtaining datasets used for training and testing which remain few and an imbalance between each type of letter (there are several types with only very few data available and this means there is a need for oversampling several times), (2) the small number of epochs which was only 20, (3) Much confusion in understanding the classification for the types of letters (which was conducted by human operators) such as official memorandum letters that apparently discuss assignments, reports, introductions, announcements, and others, and (4) the hierarchical rules of the letter from the institution where an official document was observed to be a relation tree for several types of letter criteria such as the Manuscripts of the Regulatory Service which was a relation tree for circular letter, regulation letter, and procedure as well as the assignment service manuscript which was a relation tree for instruction letter, warrant, and letter of assignment. This means it is highly likely for different types of letters to have similar contents and this makes the results obtained in classifying the 22 types of letters using the CNN method not to be optimal. Therefore, there is a need for further research and trials to achieve better classification results for the types of letters.

The regex pattern adapted from [29] to classify 15 origins of the letter and 25 subjects of the letter was evaluated and found to be accurate and precise with 100% accuracy and the same trend was also observed for the one modified to classify 5 manuscripts from the 4 used in the previous research [29]. This means the automatic hierarchical classification model is capable of classifying documents based on a hierarchy and can also be used to automatically classify during the archiving stage as indicated by the findings of this research. However, the weakness of this model is that the characteristics of the document have a significant effect on the classification results for both the types CNN and the origin of the letter using Regex. The presence of an error in the classification of the type of letter usually affects those related to the manuscript due to the relationship between the concepts.

Table 2 shows the invalid classification results for 100 digital correspondence documents and the characteristics of the document were found to have a significant effect on the results of the classification for the origin and subject of the letter. This is because letter-number has information indicating the code for the origin and the subject of the letter and this is very likely to produce an error in the classification results, specifically when the number is handwritten which is against the standards and coding rules that require dots or spaces and when there is more than one letter-number information in a document such as those with a reference letter-number. The regex pattern in the automatic hierarchical classification model is considered effective when the letter-number information can be read clearly and correctly, follows the proper writing standards, and only one is present in a document. However, the regex pattern was unable to detect letter-number since the number used for the classification is presented at the top/first.

Table 2. Invalid results of automatic hierarchical classification using CNN+Regex (from 100 documents)

No	Error type	Cause	Total of document
1	Invalid classification result of type of letter CNN	Misclassified	1
2	Invalid classification result origin and subject of letter (Regex) unsolved error	The system is limited to the matching of strings according to the specified Regex pattern but unable to detect the presence of the letter-number used at the top/first. There may be an error in the classification results for the documents with several letter-numbers	5
3	The error is not caused by the classification model. Error due to data not being properly recognized during extraction with OCR.	The characters in the letter-number are illegible because they are handwritten and this led to the failure in classifying the origin and subject of the letter. Sometimes, the regex pattern matches another string in the document that matches the pattern but not what it was intended (unsolved error explain above (No.2))	5
4	The error was not caused by the classification model. Error due to human error because the writing on the letter is not in line with the applicable standards	The data does not match the information provided in the regular expression pattern.	10

Table 3 shows the findings from the evaluation of automatic hierarchical classification of 100 documents and it was discovered that 79 were classified accurately based on the type of letter, text, origin, and subject while 21 documents were not classified accurately, thereby indicating the accuracy was 79%. The detail of those not accurate includes 1 document which is related to the misclassification of letter type CNN and 20 related to the origin and subject (Regex) which was associated with several factors such as the presence of 5 documents with letter-number handwritten which could not be extracted through an OCR. There are also 10 documents with letter-number that do not follow the coding standards or rules set and this

made it impossible for the Regex pattern to determine the strings correctly. It was also discovered that 5 documents have errors that cannot be handled in the process. Moreover, the presence of more than one letter-number in 1 document interfered with or confused the regular expression pattern which was used to retrieve the matching string. For example, when the letter-number at the top of the document is handwritten or written using inappropriate standards and the existence of other letter-numbers which can be read clearly and meet the existing Regex pattern can lead to the classification based on the criteria but the results will be incorrect. There is no special method to handle this kind of error and no special conditions have been made to select the letter-number at the top of the document like the one to be used for the classification process. Each digital correspondence document was uploaded to the system to be archived after which the extraction process to retrieve text was conducted using Tesseract OCR while the automatic classification process was conducted using the automatic hierarchical classification model developed to obtain the document criteria to classify the manuscripts, type, origins, and subject of the letter. The archiving system was used to generate the classified documents through the extraction of the text content and the results of the automatic classification are stored in the PostgreSQL database.

Table 3. Evaluation of automatic hierarchical classification (100 documents)

No	Description	Number of documents
1	Accurately classified documents	79
2	Documents not classified accurately	21
	Total documents	100
	Level of accuracy	$79/100 \times 100\% = 79\%$

Table 4 shows the archiving results for 100 documents and all the documents were discovered to have 100% extracted with 95% accuracy of the recognized text with a slight error due to the poor quality of the scanned documents which are quite blurry. Moreover, the automatic classification process for the manuscript, type, origin, and subject of the letter criteria was observed to be 79% successful for all documents. It was discovered that the application of the CNN method for the types of letters produced 99% accuracy while the Regular Expression method applied to the origin, subject, and manuscript of the letter had 80%. All document information including the id, document name, content, classification results based on the manuscript, type, origin, and subject of the 100 documents were also successfully stored with 100% accuracy in the PostgreSQL database.

The search method developed was able to improve the quality of the search results for the digital correspondence documents as indicated by the more specificity, accuracy, and fastness when compared to the use of document naming or annotations and unclassified content in the search process. It is important to note that the search was conducted using the document name (default), document content, and classification criteria to display the relevant digital correspondence documents with their ontology relationships. The improvement in the quality of the content-based search for the digital correspondence document was achieved through the availability of the required criteria such as the classification and display of the ontology relationships information to ease the users' understanding of the hierarchy of the letter found and also to display the required document. This facility is unavailable in conventional search which is designed based on the document name or annotations [3] as well as unclassified content [10]-[15].

Table 4. Archiving result (100 documents)

No	Process	Success of process	
		Percentage (100 Doc)	Accuracy
1	Extraction Text	100 %	95%
2	Automatic Hierarchical Classification	100 %	79%
3	Storing of Information Database	100 %	100 %

A trial was conducted as an example by searching for document names through the input of the query "bantuan pemerintah" and no document was shown and this complicated the search process. Figure 2 shows that the documents from content-based document search results can be selected to be displayed. These documents do have the required content, namely "*Bantuan Pemerintah*" (in the Indonesian Language). The application of the conventional search based on content produced 6 documents including 152668_Scan007 (Figure 2(a)), 145446_Peraturan Juknis K13_OK (Figure 2(b)), 739227_1416.C5.1.KU.2018 (Figure 2(c)), 435300_PERATURAN KPA INKLUSIF (Figure 2(d)), 582718_Scan0285 (Figure 2(e)), and 755995_1490.C5.3.2.KU.2018 (Figure 2(f)) which were not previously found using only the document name.

However, these results need to be examined further to determine the information on the manuscript, type, origin, and subject of the letter from the documents received. The method proposed in this research was also applied to conduct the search and 6 documents with such content were successfully produced in addition to the information on the classification ontology.

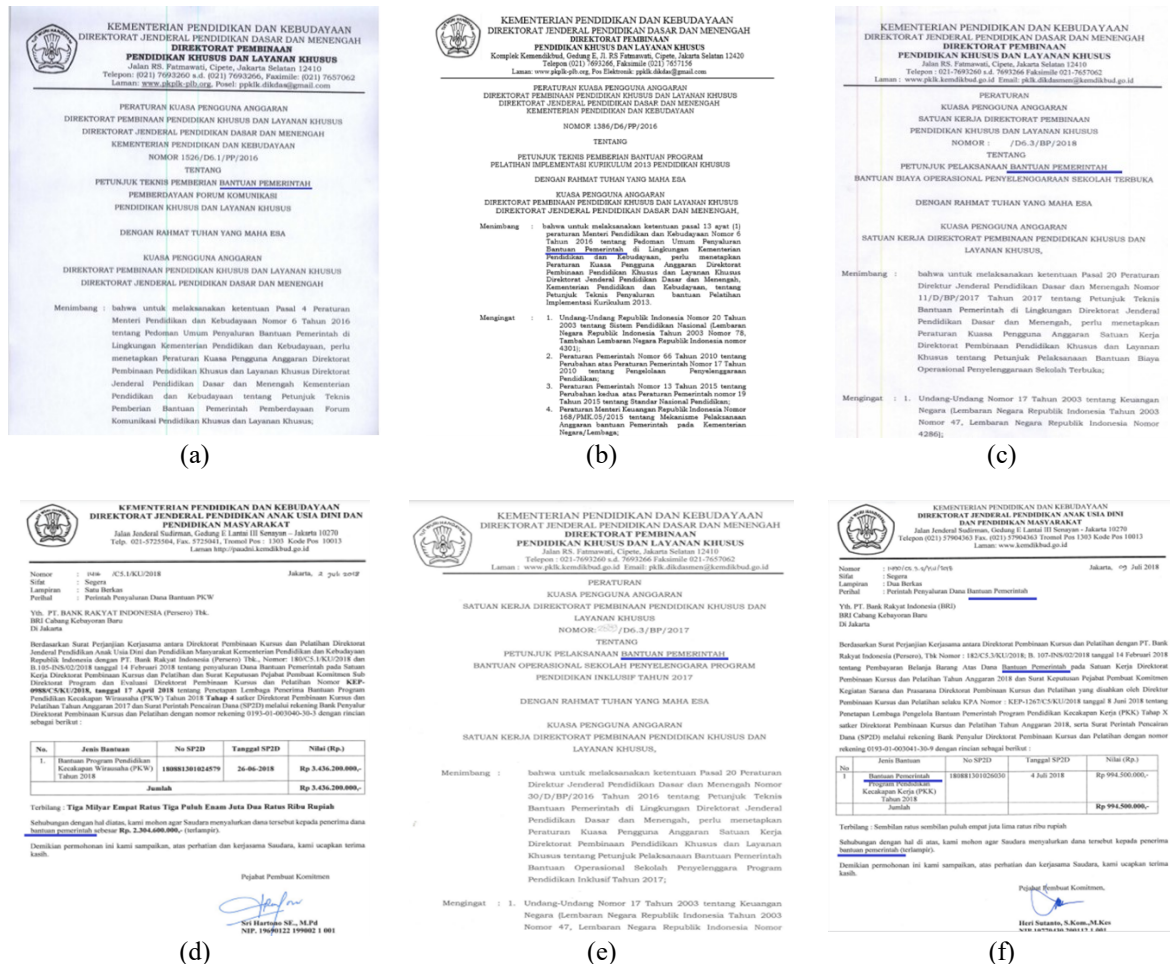


Figure 2. The documents from content-based document search results are displayed query: “*bantuan pemerintah*” (government assistance) (in the Indonesian language) for (a) 152668_Scan007, (b) 145446_Peraturan Juknis K13_OK, (c) 739227_1416.C5.1.KU.2018, (d) 435300_PERATURAN KPA INKLUSIF, (e) 582718_Scan0285, and (f) 755995_1490.C5.3.2.KU.2018

Figure 3 shows the ontology information on the hierarchical classification of content-based document search results based on query: “*bantuan pemerintah*” (government assistance) (in the Indonesian Language) of the manuscript of the letter, the type of letter, the origin of the letter, and the subject of the letter from these documents specifically, quickly, and easily. Table 5 shows the trials for the search of digital correspondence documents based on the classified content through the application of automatic ontology relationships conducted using the document name (default), content or classification criteria. Each query process formed an automatic ontology graph that shows the classification information of the required document. This search method produced highly accurate results of 100% which was obtained based on the comparison of the number of documents according to the question with the number of appropriate documents in the database multiplied by 100%. Moreover, the information on the relationships of the document was displayed to assist the users in understanding the required documents.

The overall results of the trials showed that the automatic hierarchical classification, archiving of the digital correspondence documents with automatic classification, and searching for the documents based on the classified content to display the ontology relation information were successfully conducted. This means the advanced content-based retrieval of digital correspondence documents with ontology classification was able to increase the efficiency and quality of digital correspondence documents search results more

Advanced content-based retrieval for digital correspondence documents ... (Rifiana Arief)

specifically, relevantly, accurately, and quickly than the conventional method which involves searching documents based on name or annotations and unclassified contents.

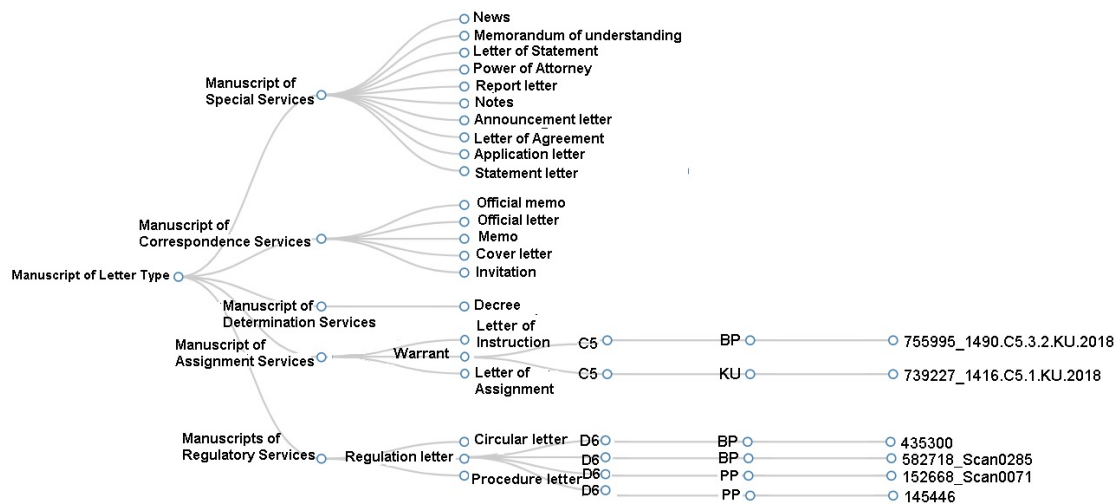


Figure 3. Ontology information on the hierarchical classification of content-based document search results query: “*bantuan pemerintah*” (government assistance) (in the Indonesian language)

Table 5. Retrieval results for the advanced content-based retrieval system for digital correspondence document with ontology classification

No	Criteria	Query	Number of docs found	Classification result with graph ontology	Retrieval accuracy
1	Name of doc	<i>Bantuan pemerintah</i>	0	is appropriate	100%
2	Content of doc	<i>Bantuan pemerintah</i>	6	is appropriate	100%
3	Classification criteria (type of letter)	<i>Surat peraturan</i>	5	is appropriate	100%
	Classification criteria (origin of letter)	D6	2	is appropriate	100%
	Classification criteria (subject of Letter)	KU	7	is appropriate	100%

4. CONCLUSION

The advanced content-based retrieval system for digital correspondence documents with ontology classification (OCR-assisted) provided a solution to the archiving problems associated with the manual systems applied in conventional digital document search using names or annotations as well as those designed to conduct content-based search assisted by OCR which is unable to classify and provide information on the ontology relationships. The contribution of this research has two novelties compared to several previous research. The first is the development of an automatic hierarchical classification model (5 manuscripts of letter \Rightarrow 22 types of letter \Rightarrow 15 origins of letter \Rightarrow 25 subjects of the letter) which can be used to archive documents in the database according to their hierarchy. The second is the advanced content-based search that identifies digital correspondence with specific text content and it is also possible to perform a more specific search based on the document classification criteria (based on the letter's type, origin, or subject), and to display the ontology of found documents. The model was observed to have succeeded in improving the quality of document search by making it more specific, accurate, and quicker than searching based on the document name or annotations and unclassified contents. This was indicated by the ability of the archiving method to perform automatic stratified classification according to the hierarchy by merging CNN and regular expression with 79% such that the accuracy of the classification using CNN was recorded to be 99% while those conducted using Regex was 80%. It is also important to note that a classification trial was conducted using only 100 documents that do not represent all types of letters with 99 classified correctly through the CNN based on the selected criteria. Meanwhile, classification using Regex produced accurate but non-optimal results due to some errors identified in the documents such as handwritten letter-numbering, numbering without following predetermined standard coding rules, and having more than 1 letter-number such as the existence of a reference number. This indicates it is necessary to create an additional condition to

anticipate the possibility of several strings in one document and select the first that fits with the pattern to improve the accuracy of the classification results using regular expressions. It is important to note that the letter-number normally used to classify the origin and subject of the letter is the number at the top of the document. Moreover, the results showed that the accuracy of searching classified content-based documents by displaying the ontology relation was 100%. It is recommended that the classification of documents with poor scan results, unstructured document text specifications, and similarity between the content of one letter and another using the CNN should be examined further. In the future, archiving methods at different institutions that implement automatic hierarchical classification by adjusting the applicable correspondence rules can also be further developed for the required search for digital correspondence documents based on the classified content with specific, accurate, and fast ontology classifications.

ACKNOWLEDGEMENTS

The authors appreciate the Directorate of Research and Community Service Directorate General of Research and Development of the Ministry of Research, Technology and Higher Education, Republic Indonesia for funding leading college applied research and General Bureau of the Secretariat General of the Ministry of Education and Culture, Republic Indonesia for allowing the use of non-confidential digital correspondence documents.





REFERENCES

- [1] A. M. Riad, H. K. Elminir and S. Abd-Elghany, "A literature review of image retrieval based on semantic concept," *International Journal of Computer Applications*, vol. 40, no. 11, pp. 12-19, 2012, doi: 10.5120/5008-7327.
- [2] P. K. Bhagat and P. Choudhary, "Image annotation: Then and now," *Image and Vision Computing*, vol. 80, pp. 1-23, 2018, doi: 10.1016/j.imavis.2018.09.017.
- [3] C. L. Philip Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314-347, 2014, doi: 10.1016/j.ins.2014.01.015.
- [4] P. Bottoni, et al., "Facilitating interaction and retrieval for annotated documents," *International Journal of Computational Science and Engineering*, vol. 5, no. 34, pp. 197-206, 2010, doi: 10.1504/IJCSE.2010.037675.
- [5] S. Kutade and P. Dhamal, "Efficient document retrieval using annotation, searching and ranking," *International Journal of Computer Applications*, vol. 108, no. 5, pp. 1-3, Dec. 2014, doi: 10.5120/18904-0198.
- [6] W. Yu and J. Hsu, "Content-based text mining technique for retrieval of CAD documents", *Automation in Construction*, vol. 31, pp. 65-74, May 2013, doi: 10.1016/j.autcon.2012.11.037.
- [7] F. Alaei, A. Alaei, M. Blumenstein, and U. Pal, "A brief review of document image retrieval methods: Recent advances," *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 3500-3507, Jul. 2016, doi: 10.1109/IJCNN.2016.7727648.
- [8] A. N. Bhute and B. B. Meshram, "Text based approach for indexing and retrieval of image and video: A review," *Social Science Research Network Electronic Journal*, 2014, doi: 10.2139/SSRN.3430312.
- [9] M. Chawla, R. Jain, and P. Nagrath, "Implementation of tesseract algorithm to extract text from different images," *Social Science Research Network Electronic Journal*, 2020, doi: 10.2139/SSRN.3589972.
- [10] N. Ramadijanti, A. Basuki and G. J. W. Agrippina, "Designing mobile application for retrieving book information using optical character recognition," *2016 International Conference on Knowledge Creation and Intelligent Computing (KCIC)*, 2016, pp. 176-181, doi: 10.1109/KCIC.2016.7883643.
- [11] V. Aggarwal, S. Jajoria, and A. Sood, "Text retrieval from scanned forms using optical character recognition," *Sensors and Image Processing*, pp. 207-216, Oct. 2017, doi: 10.1007/978-981-10-6614-6_21.
- [12] J. M. Jayoma, E. S. Moyon and E. M. O. Morales, "OCR based document archiving and indexing using PyTesseract: A record management system for dswd caraga, Philippines," *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2020, pp. 1-6, doi: 10.1109/HNICEM51456.2020.9400000.
- [13] P. A. Wankhede and S. W. Mohod, "A different image content-based retrievals using OCR techniques," *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, 2017 pp. 155-161, doi: 10.1109/ICECA.2017.8212785.
- [14] A. Chiatti, et al., "Text extraction and retrieval from smartphone screenshots: building a repository for life in media," *In Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC '18)*, pp. 948-955. Apr. 2018. doi: 10.1145/3167132.3167236.
- [15] C. Adjetej and K. S. Adu-Manu, "Content-based image retrieval using Tesseract OCR engine and levenshtein algorithm," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 7, 2021, doi: 10.14569/IJACSA.2021.0120776.
- [16] R. Bhagat, P. Thosani, N. Shah and R. Shankarmani, "Complex document classification and integration with indexing," *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1477-1484, 2021, doi: 10.1109/ICESC51422.2021.9532737.
- [17] M. Rivest, E. Vignola-Gagné, and É. Archambault, "Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling," *PLoS ONE*, vol. 16, no. 5, 2021, doi: 10.1371/journal.pone.0251493.
- [18] Md. Z. Hasan, S. Hossain, Md. A. Rizvee and Md. S. Rana, "Content based document classification using soft cosine measure," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 4, 2019. doi: 10.14569/IJACSA.2019.0100464.
- [19] M. N. Krasnyanskiy, A. D. Obukhov, and E. M. Solomatina, "The algorithm of document classification of research and education institution using machine learning methods," *2019 International Science and Technology Conference "EastConf"*, pp. 1-6, 2019, doi: 10.1109/EastConf.2019.8725319.





- [20] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, Feb. 2015, doi: 10.1186/s40537-014-0007-7.
- [21] K. Kowsari, D. E. Brown, M. Heidarysafa, K. Jafari Meimandi, M. S. Gerber and L. E. Barnes, "HDLTex: hierarchical deep learning for text classification," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 364-371, doi: 10.1109/ICMLA.2017.0-134.
- [22] H. B. Dogru, S. Tilki, A. Jamil, and A. Ali Hameed, "Deep learning-based classification of news texts using Doc2Vec model," *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pp. 91-96, Apr. 2021, doi: 10.1109/CAIDA51941.2021.9425290.
- [23] M. Ali Ramdhani, D. S. Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 19, no. 2, pp. 1000-1009, Aug. 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.
- [24] X. Sun, Y. Li, H. Kang, and Y. Shen, "Automatic document classification using convolutional neural network," *Journal of Physics: Conference Series*, vol. 1176, no.3, pp. 032029, Mar. 2019, doi: 10.1088/1742-6596/1176/3/032029.
- [25] A. Kolsch, M. Z. Afzal, M. Ebbecke, and M. Liwicki, "Real-time document image classification using deep CNN and extreme learning machines," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1318-1323, Nov. 2017, doi: 10.1109/ICDAR.2017.217.
- [26] B. Hassan, R. Amina, L. Amine, L. Elhoussin, and M. Azouazi, "A regexcriteria API to complete the power of regular expressions engine," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 3185-3193, Aug. 2019, doi: 10.11591/ijece.v9i4.pp3185-3193.
- [27] D. D. A. Bui and Q. Zeng-Treitler, "Learning regular expressions for clinical text classification," *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 850-857, Sep. 2014, doi: 10.1136/AMIAJNL-2013-002411.
- [28] M. Cui, R. Bai, Z. Lu, X. Li, U. Aickelin, and P. Ge, "Regular expression based medical text classification using constructive heuristic approach," *IEEE Access*, vol. 7, pp. 147892-147904, 2019, doi: 10.1109/ACCESS.2019.2946622.
- [29] R. Arief, A. B. Mutiara, T. M. Kusuma, and H. Hustinawaty, "Automated hierarchical classification of scanned documents using convolutional neural network and regular expression," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 1, p. 1018, 2021, doi: 10.11591/ijece.v12i1.pp1018-1029.
- [30] E. Aroua and A. Mourad, "An ontology-based framework for enhancing personalized content and retrieval information," *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, pp. 276-285, 2017, doi: 10.1109/RCIS.2017.7956547.
- [31] J. Wang, J. Liu, and L. Kong, "Ontology construction based on deep learning," *Advances in Computer Science and Ubiquitous Computing*, pp 505-510, 2018, doi: 10.1007/978-981-10-7605-3_83.
- [32] K. Nyberg, T. Raiko, T. Tiininen, and E. Hyvönen, "Document classification utilising ontologies and relations between documents," *Proceedings of the Eighth Workshop on Mining and Learning with Graphs - MLG '10*, 2010, doi: 10.1145/1830252.1830264.
- [33] J. C. Rendon-Miranda, J. Y. Arana-Llanes, J. G. Gonzalez-Serna, and N. Gonzalez-Franco, "Automatic classification of scientific papers in PDF for populating ontologies," *2014 International Conference on Computational Science and Computational Intelligence*, Mar. 2014, doi: 10.1109/CSCI.2014.153.
- [34] N. Sanchez-Pi, L. Martí, and A. C. B. Garcia, "Improving ontology-based text classification: An occupational health and security application," *Journal of Applied Logic*, vol. 17, pp. 48-58, Sep. 2016, doi: 10.1016/j.jal.2015.09.008.

BIOGRAPHIES OF AUTHORS







Rifiana Arief     Lecture at the Faculty of Computer Science and Information Technology, Gunadarma University. Deputy Head of Computer Network Development Laboratory. Her research interest includes artificial intelligence, big data, and computer science. She can be contacted at email: rifiana@staff.gunadarma.ac.id.







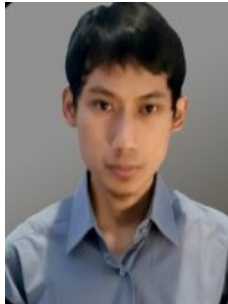
Suryarini Widodo     Lecture at the Faculty of Computer Science and Information Technology, Gunadarma University. Head of Internet Development Laboratory. Her research interest includes handwriting recognition and image processing. She can be contacted at email: srini@staff.gunadarma.ac.id.







Ary Bima Kurniawan     Lecturer at the Faculty of Computer Science and Information Technology, Gunadarma University. Staff of Computer Network Development Laboratory. His research interest includes computer networking, web programming, IoT, and blockchain. He can be contacted at email: bima@staff.gunadarma.ac.id.



Hustinawaty     Lecturer at the Faculty of Computer Science and Information Technology, Head of Programme Magister Information Management, Gunadarma University. Her research interests are image processing. She can be contacted at email: hustina@staff.gunadarma.ac.id.



Faisal Arkan     Student at Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia. His research interests are web programming and artificial intelligence. He can be contacted at email: faisalarkan21@gmail.com.