https://doi.org/10.5392/IJoC.2017.13.4.070

Patent Document Similarity Based on Image Analysis Using the SIFT-Algorithm and OCR-Text

Jeong Beom Park
Department of Information and Telecommunication
Pai Chai University, Daejeon, 155-40, Republic of Korea
University of Hildesheim, Hildesheim, 31141, Germany

Thomas Mandl
University of Hildesheim, Hildesheim, 31141, Germany

Do Wan Kim
Pai Chai University, Deajeon, 155-40, Republic of Korea

ABSTRACT

Images are an important element in patents and many experts use images to analyze a patent or to check differences between patents. However, there is little research on image analysis for patents partly because image processing is an advanced technology and typically patent images consist of visual parts as well as of text and numbers. This study suggests two methods for using image processing; the Scale Invariant Feature Transform(SIFT) algorithm and Optical Character Recognition(OCR). The first method which works with SIFT uses image feature points. Through feature matching, it can be applied to calculate the similarity between documents containing these images. And in the second method, OCR is used to extract text from the images. By using numbers which are extracted from an image, it is possible to extract the corresponding related text within the text passages. Subsequently, document similarity can be calculated based on the extracted text. Through comparing the suggested methods and an existing method based only on text for calculating the similarity, the feasibility is achieved. Additionally, the correlation between both the similarity measures is low which shows that they capture different aspects of the patent content.

Key words*: Patent Similarity, Image Processing, Information Retrieval, Correlation Coefficient, SIFT, OpenIMAJ, OCR, Tess4j.*

1. INTRODUCTION

The amount of technological information available is increasing rapidly and the economic value of information has become an important factor. In this situation, users need to find relevant information that they want and Information Retrieval (IR) has become one of the major technologies in Information Technology (IT).

In this situation, various technologies have been developed and the technologies have been shared easily through the Internet. It has become one of the most important elements in the information society to protect Intellectual Property (IP) which means protecting the right of the creative work of the inventor. There are several legal mechanisms to ensure IP rights and one of them is the patent. According to the World Intellectual Property Organizatio [1], a patent is an exclusive right for exploiting an invention which is a new solution or a new method for a problem. Additionally, it includes various technologies in most industries and fields. Therefore, it is necessary to regard patents as a source of information which can indicate new technologies, trends and the direction of research through automatic analysis. Due of these reasons, IR and text mining are important technologies in the patent area. To put it simply, IR is used to search related studies or to check for similar inventions. This helps to avoid double investment for developing new technologies. Furthermore, retrieval technology can be used to search for relations between patents and identify e.g. patent landscapes. Additionally, the content of patents is analysed in order to make it possible to retrieve them. In patent documents, there are different elements such as description, claim, image, IPC[1], document number and so on. According to the article [2], patent users consider images in patents as an important element for patent searching. In particular, the patent image is one of the most regarded sections

* *Corresponding author, Email: dwkim@pcu.ac.kr*
Manuscript received Oct. 20, 2017; revised Dec. 18, 2017; accepted Dec. 18, 2017

[1] International Patent Classification

www.kci.go.kr

during the results assessment at least for many technological areas. Besides, the patent users opinion is that image search is one of the most desired functions for future patent search systems. According to the article [3], when patent experts are searching for a patent through a search system, they can distinguish a related patent document by images in the list for searching the document. It is possible to save a lot of time and useful for searching the document. Because of this reason, there are studies about classifying the patent image for each type [4], [5]. Lastly, the article [6] shows that images are one of the most important elements for trend mining in the patent through interview with information professionals working in the patent domain. Therefore, patent images are an important element in the patent document and it is possible to use images to analyse the patent.

However, image processing technology is hard to use difficult technology within IT and requires much processing power. So, most research on patent retrieval has been dedicated to text because text technology is more mature and easier available than image processing. There are a few research reports on patent retrieval using image processing [4], [5], [7]. The main objective of the study is to suggest two methods based on the image analysis for patent retrieval and then to calculate the document similarity between the suggested method and an existing method. If the result shows a difference in document similarity, it could be used as a new method for the patent retrieval. The new methods based on image analysis are shown below:

- **SIFT-Algorithm**: The analysis based on image processing technology is applied to the patent images
- **OCR-Text**: This part of the analysis is based on text extracted from the patent images

The first method works with the SIFT algorithm to analyse the patent document by using only the images. It uses the image feature points to calculate similarity between documents. And, the second method works with the OCR technology to analyse the patent document by using the text which is extracted from an image. It uses the same method as the existing text based analysis after extracting text from the images. When the results of these methods are compared, it could be shown how different the results of the new methods and the existing method are. Our study uses the document similarity to compare them.

2. RELATED WORK

2.1 Text Preprocessing

2.1.1 Preprocessing: Preprocessing is a method to improve the effectiveness of text processing. There are many methods for preprocessing. This subsection will provide detailed explanation about the basic and most important elements of text operations. The article [8] mentioned three basic tasks: Tokenization', 'Stop word elimination' and 'Stemming'. Tokenization divides the string instances of a text into the smallest possible meaningful units. Sentences are divided into words. Usually, it is carried out by using spaces and other delimiters between words to separate units. There are some words which do not carry significance for a document within the context of information retrieval even if they appear frequently. These words are called 'stop words' and should be removed. Usually, a stop word list is used to remove them. This language specific list consists of articles, prepositions, pronouns, be-verbs conjunctions and so on. The last processing step is stemming. There are many different forms of the same word in a document such as 'play', 'plays' and 'playing'. Even if they have the same meaning, their grammatical form is different. By applying stemming, it is possible to reduce the words to their stem and therefore to classify the different forms of a word accordingly [9].

2.1.2 Weighting

Simple pattern matching and Boolean operations do not lead to good results in search. Instead, results should be presented as a ranked list according to their degree of relevance. Thus, it is necessary to find information for the ranking. Term weighting has been used to resolve this problem and there are three main components [10].

Term Frequency (TF) is to give weights by term frequency in document. That is, if a word appears more often in a document, it seems to represent the content of that document better. So, the document weightings by TF seem to represent values of the document. This is referred to as the bag of words approach because only occurrences of words are considered not their order [9].

The Inverse Document Frequency (IDF) is used to reduce the weight of a word which has a high frequency in all documents. According to Zipf's law [11], words which appear frequently could be less important for content representation in information retrieval. So, it is necessary to identify the importance of a specific word in all documents. The formula below shows methods how to calculate IDF. N is the number of total documents and Document Frequency (DF) is the number of documents which has a specific word(t). Also, one is added for preventing a division by zero(0).

$$\mathrm{IDF(t, N)} = \log \frac{|N|}{|1 + DF_t|} \tag{1}$$

All documents have a different number of words. When calculating the weights without document normalization, it could cause a problem. For example, the first document has a total 100 words and the second document has a total 200 words. When searching a specific word, the first document has the 20 words and the second document has the 25 words. Even if the second document includes the specific words more frequently than the first document, the first document shows a higher frequency of the specific word. The document normalization resolves this problem and provides the same conditions to all documents in order to calculate weighting.

2.1.3 Vector Space Model

Vector space model is the model in which a set of documents is represented as vectors in the vector space. It is

www.kci.go.kr

used for various information retrieval calculations such as document similarity, classification, clustering, indexing and so on [9]. A document is represented as a vector and each word is represented as one dimension. Therefore, if a document has a specific word, the value of the corresponding dimension about the specific word is not zero(0). The value of dimension is the term weight in a document. Usually, TF-IDF method is used for calculating the weighting of each word in a document.

2.2 SIFT-Algorithm

The Scale Invariant Feature Transform(SIFT) algorithm is one of the most famous algorithm to extract features from an image. The most important characteristic of SIFT is that it is possible to extract invariant feature points without impacting size, direction, noise and change of image. This character is an important point to understand the basic logic of SIFT. In the SIFT algorithm, there are two big steps. One is feature extraction, the other one is feature matching. The feature extraction is to extract feature points from an image and the feature matching is to find matching points between the images.

2.2.1 Feature Exraction

This step is to extract feature points from an image. There are four tasks in this step.

- **Scale-Space Extrema Detection**
- **Keypoint Localization**
- **Orientation Assignment**
- **Keypoint Descriptor**

The Scale-Space Extrema Detection task is to extract candidate points which could be the feature point. There are three steps in this task. One is to calculate the Gaussian pyramid [12], the other one is to calculate the Different of Gaussian(DoG) [13] and the last one to calculate the candidate points. The image pyramid is used for a character of a scale invariant and the Gaussian pyramid is one method to make the image pyramid. It has advantages to create image pyramid fast and it only needs small memory capacity. Gaussian image is created by applying the Gaussian filter which has a variance to an image through convolution. The DoG is calculated by different of these Gaussian images. The number of the DoG are decided as much number of Gaussian image minus one. The last step is necessary to have three adjacent DoGs to calculate the points. When comparing the query point and the target points, if the query point has an extreme point which are the highest or lowest from all target points, the query point is extracted as the candidate point

The Keypoint Localization task is to find the feature point from the candidate points. In this task, there are two steps. One is to remove a low contrast key point and the other one is to remove an edge point. Taylor expansion [14] and Harris corner detector algorithm [15] are used in each step.

The Orientation Assignment task is to calculate direction and size of the feature point in order to find a matching point between images which has a difference direction. 16×16 pixels, which are nearby a feature point, are used to calculate the value of direction and size of the feature point. In this task, if there is a value which is over 80% of the maximum value, it can be another the orientation value of the feature point. That is, it is possible that one feature point has several orientation values.

Through previous tasks, an image location, scale and orientation to a feature point are calculated. These parameters are for 2D coordinate system. The Keypoint Descriptor task is to calculate a vector of the feature point in order to have an invariant character in illumination or 3D-viewpoint. The normalization of orientation is used for the illumination invariant. The 16×16 gradient are represented by the 4×4 gradient and it has 8 directions. Therefore, the descriptor vector has 128-dimensions.

2.2.2 Feature Matching

The feature matching is to find matching feature points between two images. Usually, a query point means a input feature point and a target point means a comparison target. The Euclidean distance is used to calculate the matching point. The minimum value of distance is the matching point. According to the article [13], there were many wrong cases when finding the matching point by using only the distance. So, a method is used to calculate the matching point. The method is that if the rate between the first minimum distance and the second minimum distance is over 0.8, the feature point can't use as the matching point. Therefore, the result of matching points depends on the query point due to calculating relatively.

2.3 OCR-Text

Optical Character Recognition(OCR) is one of the pattern recognition technologies and it began from the method to recognize a character by the pattern matching. It is a technology that a image, which is painted by human or computer, changes to a code that a computer can recognize. According to the article [16], OCR is used in various fields such as Invoice imaging, Banking, Optical Music Recognition(OMR), CAPTCHA[2] and so on. In particular, OCR is often used to extract text from a image. There many OCR programs based on the open-source. Tesseract [17] is the most popular engine and has the powerful performance to extract text from image. Hewlett-Packard(HP) developed the Tesseract in 1984. The Tesseract has been upgraded consistently and HP released the open-source based the Tesseract in 2005. Now, Google is supporting some part of the Tesseract.

There are four steps in the Tesseract. The first step is to make a binary image by using a thresholding of image. This step is to change an image to a code that computer can recognize. The second step is to analyse connected components of the binary image and to extract outline of the components. The connected components mean the number of component to identify the shape of one letter. For example, Letters 'i' and 'j' has two components and letters 'a' and 'b' has one component. Because of this step, Korean is harder than English to extract the text from the image. The third step is to systematize the components by analysing a text line and then the text is divided as a unit of word by space between letters. The last step is to recognize the divided word by the unit of word or page.

[2] Completely Automated Public Turing test to tell Computers and Humans Apart

www.kci.go.kr

2.4 Image Processing in Patents

So far, there exist only a limited number of approaches for including visual information in patent search. One examples is the similarity search in the system PatMedia [18]. First experimental systems for concept based graphic search have been developed [19]-[21].

SIFT-like local orientation histograms have been applied to represent images from patents in Fisher-Vectors [4]. Within a shared task in the framework of the CLEF-IP evaluation, the best retrieval results were obtained by integrating textual and visual information in late fusion approaches [3]. An overview for the classification of types of images (photograph, flow-chart, technical sketch, diagram, graph) is provided by the article [22]. Given the heterogeneity of images in patents, this research is also highly relevant for patent information retrieval.

3. PROPOSED ALGORITHM

3.1 Data Collection

The main objective is to collect text and image data. In particular, the collection of separate images separated image in order to apply the image processing techniques. Usually, almost every patent office provides textual information about patent documents. On the other hand, it is difficult to find a patent office that provides patent image data online. This part focuses on how to collect image data.

Two databases are used to collect text and image data:

- **WIPO Database**: To collect XML which has text and image information
- **Google Image Database**: To collect images by using information in XML

There are two processing steps necessary. One is to get the XML files from the WIPO database. The other one is to get the corresponding image files from the Google image database. With this method it is possible to collect image data by using image file information from the XML file obtained for each patent. WIPO provides patent documents from various countries such as US, Canada, Europe and so on. However, only the US data was used for the whole task. That is because sometimes there is no image data in the Google image database in patents from outside the US. The first step is the Web crawling to collect zip files which contain the XML files. The next step is to extract the XML file and to get the image information by using a DOM parser [23]. The last step is to collect images by using the URL which combines the Google database URL and the image information.

For this experiment, the query ‘Contact Lens’ was used. The article [24] shows the trend of contact lenses in the patent field from 1994 to 2014. Even if a contact lens is considered a medical device, many people have used the lens for various purposes such as to change the color of their eyes, to avoid glasses and so on. So, research on contact lenses has been rapidly growing since 2010. Documents are collected as XML files from the WIPO Web site and images are collected from the Google image database. Table 1 shows the result of the query.

Table 1. Number of documents and images by query

Total Documents	Total Images
1,116	9,993

3.2 Text Processing

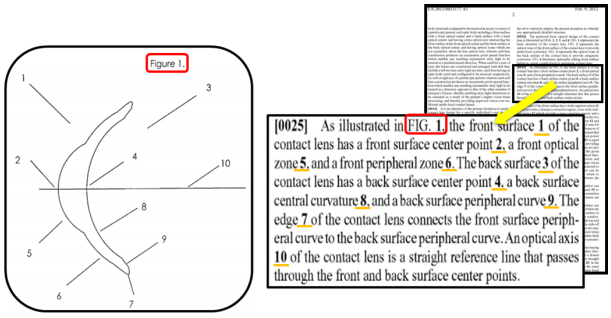

Fig. 1. The relationship between image and text in a document [25]

As can be seen in Fig. 1, this section deals with the processing to extract texts which has image information from the description part in the documents before the indexing task. It is one of the main sources of OCR text. There is textual information about images in the description part in the document. Fig. 1 shows the image and the paragraphs in the description which explain the image in the document. In the image, each number points to a specific part through a line or an arrow. In the description of the documents, there is text which explains the meaning of numbers or how to use that part. However, it is not obligatory to explain every number and image in the description. There are three major tasks to extract the text which contains image information.

- **To collect separated figure numbers**
- **To collect numbers in each image**
- **To extract text based on the number in an image**

XML tags were used in order to carry out this task. In XML, there are three kinds of tags which identify image information and they are added to the standard XML file from WIPO [26]. The first task is to extract a figure number list by using the tag ‘figref’. This is done in order to identify how many figure numbers there are in the description. In a patent document, there are two kinds of words to represent a figure number. They are ‘Fig’ and ‘Figs’. ‘Fig’ means that there is one figure number and ‘Figs’ means that there are more than two figure numbers. In ‘Fig’, to get a separated figure number, the text of the tag can be used without being modified. On the other hand, the figure number text needs to be modified when using the ‘Figs’ tag. There are three different types which represent ‘Figs’ seen below Table 2.

Table 2. Three types to represent ‘Figs’

Type	Existing Figure	Separated Figure
1	Figs. 1 and 2	Fig.1, Fig.2
2	Figs. 1 to(-) 3	Fig.1, Fig.2, Fig.3
3	Figs. 1, 2, and 5	Fig.1, Fig.2, Fig.3

www.kci.go.kr

The second task is collecting numbers in images by using the tag 'b'. As mentioned above, the 'b' tag includes wrong numbers. These numbers are not numbers in the image but figure numbers. Therefore, the main task is to collect 'b' tag list which is a list of numbers in an image. Also, this will be operated along with another work to find a letter which is included in the number instead of tag 'i'. The first step is to find a number in a sentence. At this time, if it finds a number appearing together with the word 'Fig.', it would ignore that number and find the next number until the number doesn't involve the word 'Fig.'. If it finds a number in a sentence, the next task would be to gradually compare the numbers in the 'b' list and the numbers that it found. At this time, letters and characters without numbers are removed from the number in a sentence. And then, if a number which is the same as the one in the 'b' list is found, the number in the sentence will be extracted with letters and characters that are appearing together with the number.

The last task is the text based collection of numbers in the image. The numbers in the newly created 'b' list are used to achieve that. This work to extract text by using numbers from the 'b' list is similar to the second step mentioned above. First of all, it finds a number which is the same as a number in the 'b' list in the sentence. After finding the number, the text appearing in front of the number is extracted from the beginning of the sentence. And then, the starting point is set after the number in order to extract the text appearing after the number occurrence. These processes are repeated until there is no number left in the 'b' list.

Fig. 2 shows the result of the text processing by using one patent document [25]. Fig. 2(A) shows the total list of figure number in the document and the list of numbers in image for each figure number. And, Fig. 2(B) shows the total list of numbers images in the document and the list of text appearances of these numbers for each number.

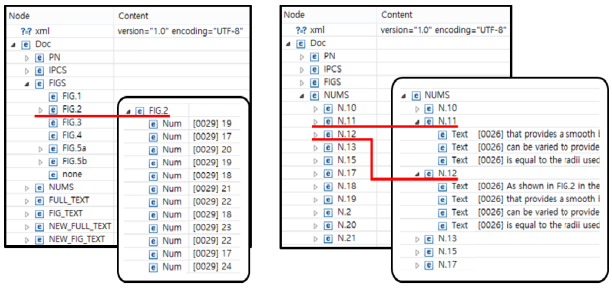

Fig. 2. The result of the text processing

3.3 Image Processing

3.3.1 OCR Text

OCR-Text is a methodology used to extract text from images. In this part, the main objective is to extract text and numbers from patent images by using OCR. Tess4j [3] is a Tessaract [17] based Java library and provides powerful OCR performance for Java applications [27]. Besides, it is possible to improve the rate of recognition of text in images by using the jTessBoxEditor[4] by training the system according to the font style [28]. Unfortunately, it is very difficult to use jTessBoxEditor for patent images because there are so many kinds of font style [5].

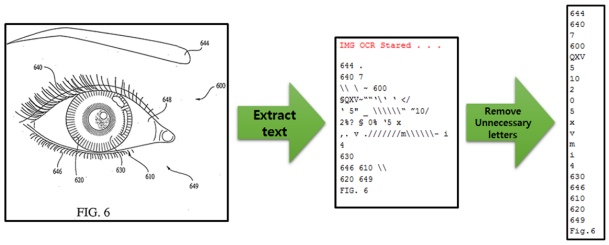

Fig. 3. The process to extract text from image [29]

Fig. 3 shows the process of extracting text from image by using Tess4j. According to the first result, there extracted letters are noise. Therefore, for the final results of the text extraction from images only text not including letters such as special symbols are used. Sometimes, a patent image contains text in other orientations. This means that some images have a 0-degree angle and others have a 90-degree angle. Therefore, in order to extract the text from all images, the algorithm was operated four times from different angles (such as 0°, 90°, 180° and 270°). Also, sometimes more than two images are concatenated on one page. This could lead to recognition problems while analyzing the images. The article [7] suggests the method to separate the images of one page with several images. However, this method can only extract a drawing without the figure numbers. Figure numbers are very important elements in image processing.

The next task is to extract text from the result of the text processing by using the extracted numbers. The result of the text processing is needed to extract text by using OCR because a patent image has very little textual information. Besides, almost all information consists of numbers. Therefore, this step suggests a new method for extracting textual information from the description by using the result of OCR and the text processing. According to Fig. 2(B), each number has some text which is associated with numbers. Therefore, it is possible to extract text through a comparison of the extracted numbers resulting from OCR and numbers resulting from text processing. Besides, an extracted figure number is also used to extract text. Fig. 2(A) shows the list of numbers extracted from figures. It is possible to extract numbers in images by using the extracted figure number. Using these numbers could resolve the problem of the missing of extracted information by using OCR. This leads to two kinds of text which are gained through the OCR operation. One is text which is in an image and the other text is the result of aligning the images and the parts within the images to the text references and subsequently to the text referring to the parts within the image.

[3] "http://tess4j.sourceforge.net/" *(last retrieved : 01.09.2017).*

[4] "http://vietocr.sourceforge.net/training.html" *(last retrieved : 01.09.2017).*

www.kci.go.kr

3.3.2 Image Feature

The SIFT algorithm has a very good performance and it is one of the most popular algorithms for extracting features from images. Therefore, this algorithm was used for this step. Besides, other research has used the algorithm for analyzing images in the patent field [4], [5], [7]. SIFT is particularly useful for applications to patent images. It extracts unique features that never change even though the image size and direction might be modified. This means that it is possible to compare images which have different sizes or directions. As mentioned above, sometimes the patent image has a different direction or two images are printed on one page. It means that even if two images are very similar, they could have different size and direction. These elements could lead to unsatisfying outcomes when comparing the images. However, using SIFT allows the comparison of images without this problem.

OpenIMAJ[5] [30] is used for the application within this task. It is a JAVA library for image processing. There are two main steps when comparing images with the SIFT algorithm. One is finding features from the images and the other one is matching these features between images. When extracting features from the images the 'DoGSIFTEngine', which is provided by the OpenIMAJ, is used. A detailed description of its implementation is given in by the article [31]. However, there is a disadvantage using this algorithm. It is too slow compared to other feature algorithms. In order to match features, the task to finding features needs to be carried out numerous times. The article [7] suggests a feature database based on image features for an image search service. This method first extracts the features from the images and then computes the feature data. The feature data consists of five parameters (such as X-axis, Y-axis, scale, orientation and 128-dimension image vector). When entering an input image, features from the image are extracted and compared with existing features in the feature database. By using this method, it is possible to reduce repeated computing processing when calculating the matching features.

After finishing the feature matching, the algorithm is used to calculate the rank by document similarity based on image features. Before calculating the rank, it has to calculate the document similarity first. Dice's coefficient [32] is used to calculate the similarity. When values have the collective characteristic of mathematical structures, it is possible to apply the values to Dice's coefficient. Besides, all documents have different numbers of images. So, it is necessary to normalize the similarity values. The simplest method is to use the average to compute the overall value. In the future, the maximum value of an image should be evaluated to calculate the value for a document.

$$S(a,b) = \frac{2 \cdot N_{a \cap b}}{N_a \cdot N_b} = \frac{2 \cdot feature_{a \cap b}}{feature_a \cdot feature_b} \quad (2)$$

[5] "http://openimaj.org/" *(last retrieved : 01.09.2017).*

3.4 Text Index

3.4.1 Preprocessing

When trying to make an indexing file by using only these methods for the first time, over 30,000 words as features are extracted in one vector space model. If the number of features is too high, it could possibly cause some problems such as 'Curse of Dimensionality' [33]. The article [34] emphasizes the significance of noun phrases in patent documents. Therefore, extracting noun phrases from the text could be a good method to resolve the problem of the 'curse of dimensionality'.

The article [35] suggests a method to extract noun phrases by using the POS tagger from Stanford [36] during preprocessing. There are three steps for preprocessing.

The first task is to remove unnecessary letters, which are called 'Delimiters', from the extracted text. For that, a function is applied which is called 'removeGarbage' and which is provided within the 'WordTokenizer' by WEKA[6].

The second task is to extract the noun phrases from the extracted text by using the POS tagger. The Stanford library provides different kinds of POS taggers. For this task, 'english-left3words-distsim' is used due to its fast processing speed. The POS tagger is used to extract the noun phrases based on n-grams; unigrams and bigrams. Unigrams are used to analyze text by every one letter and bigrams are used to analyze it by every two letters. According to the article [34], using the combination of both unigrams and bigrams is as features is beneficial. So, this task uses the POS tags based on the combination of unigrams and bigrams.

The last task is to use the stemming and stop word in order to remove unnecessary words. Apache Lucene [7] provides 'Snowball' [37] which is a framework for stemming. Stop word elimination is the process of removing words that do not carry any information, even if they appear many times. The University of Neuchatel[8] provides some kinds of stop word lists based on the languages of several countries. A stop word list consisting of 571 words in the English version is used.

3.4.2 Indexing

This task includes the creation of make three kinds of indexing file: 'Full text', 'Num text' and 'OCR text'. The first step is to call text data from the XML by using a DOM-parser. For this step, the document and OCR text XML files are used. The document XML includes three kinds of tags in description. One is the 'p' tag represents a sentence and the other one is 'figref' and 'b' tag to represent image information. In order to collect the full description text, the 'p' tag is used. The 'figref' and 'b' tags are used to collect only the sentence which has image information. And, it uses the OCR XML file to make the indexing file based on OCR text. The second step is to apply each collected text to the preprocessing described. After then, the last step is to make indexing files based on the three kinds of text.

[6] "http://www.cs.waikato.ac.nz/ml/weka/" *(last retrieved : 01.09.2017).*

[7] "http://lucene.apache.org/core/" *(last retrieved : 01.09.2017).*

[8] "http://members.unine.ch/jacques.savoy/clef/index.html" *(last retrieved : 01.09.2017).*

www.kci.go.kr

3.5 Vectorization

3.5.1 Weighting

TF-IDF is used for this task. After computing the value of TF-IDF of one vector, it has to be normalized based on the document length to provide the same conditions to all vectors. This means that each value of TF-IDF in one vector is divided by the total TF-IDF value of the corresponding vector.

Table 3. The number of documents and features

Name of Documents	Number of Documents	Number of Feature
Full Text	1,116	14,915
Num Text	1,116	5,609
OCR Text	1,116	22,812

3.5.2 Ranking Method

A ranking method is used to calculate the similarity between documents by using the Kullback-Leibler (KL) divergence [38]. It is a method to compute the difference between two probability distributions.

According to Table 3, the number of features in OCR text is the largest because the OCR text has more unique nouns than the Full-text and Num-text. When looking at Fig. 3, it is possible to check whether incorrect text is extracted by potential computation errors of the OCR system. Even if these words have no meaning, they are recognized as a unique noun. So, Table 4 shows that there are 16,120 features which appear once in the OCR text and these features are unnecessary and the similarity can be computed without them. That is because it is possible to calculate the KL divergence when both the p(x) and q(x) have no zero value in the corresponding feature. Therefore, the influence of these features on the result is negligible.

Table 4. The number of appearance about features

Number of appearance	Full Text	Num Text	OCR Text
1 time	3,947	1,865	16,120
2 times	949	784	2,264
3 times	312	427	964

3.6 Comparison Method

The comparison method includes the calculation of the similarity among the four results based on rank. The results are the Full-text based rank, Num-text based rank, OCR-text based rank and rank based on image feature. In order to compare them efficiently, only part of the ranking result, in particular the top 10, 15, 20 and 25 documents are selected for this method.

Before starting the comparison method, a problem has to be considered when calculating the rank by matching features. There are too many repetitions of the algorithm. When extracting the features from images, the algorithm needs to run 9,993 times which the number of total features of an image. On the other hand, when matching features, it is computed over 9,993 times for each of the 1,116 documents. As this requires too much processing power, an alternative method is suggested.

The new method is to choose some documents for applying the comparison method. When considering the computation time, it was decided to choose 50 documents by using rank based on the text. First of all, the documents are collected from the ranking top 10 about the first document in the each result. So, it is possible to collect a total of 31 documents. 30 documents are from the top 10 documents in the three results and one document is the first document. After removing duplicated documents, the number of collected documents is 25. The other 25 documents are collected randomly.

In order to compare the results based on rank, the Spearman rank correlation coefficient [39] is used. It is one of the correlation analysis methods which are based on rank. The last step compares now the ranking results.

4. EXPERIMENTAL RESULTS

The range of the result values of the Spearman correlation coefficient is -1 to 1(-1 ~ 1). Generally, there are three potential outcomes of a correlation analysis. One is a positive correlation, the other one is a negative correlation and the last one signifies no correlation. When the value is zero(0), it is uncorrelated. The article [39] suggests the correlation coefficient interpretation by range of value.

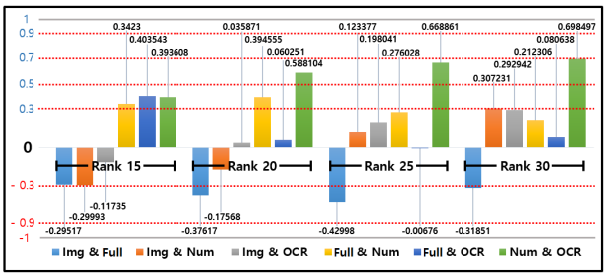

Fig. 4. The graph about correlation among results

Fig. 4 shows the correlation values between the results. According to the mentioned measurement method, negative values means that there is no similarity. Therefore, this analysis method should consider only positive values. First of all, it shows correlation values between rank based on image and text based rank through the first three data bars. On average, these values are located in -0.3 to 0.3. Even if some values, such as the first bar (blue) in rank 20, 25 and 30 as well as the second bar (orang) in rank 30, are over 0.3 or -0.3, they are negative value or close to 0.3. According to the article [32], the correlation of these values show a negligible correlation. Secondly, when extracting text data for the OCR-text, it uses Num-text. So, it shows that the sixth data bar (green) located from 0.5 to 0.7 on average. It means that the OCR-text and Num-text have a moderate positive correlation. And, the fifth bar (dark blue) are located in 0 to 0.3. This means that there is little relation between the OCR-text and the Full-text. Lastly, Num-text consists of some sentences of the Full-text. However, when comparing text length between them, the difference between them is large. Table 4 shows that the difference of the number of words between Full-text and Num-text is larger than

www.kci.go.kr

the difference of Num-text and OCR-text when removing words which appear only once. On average, the fourth bar (yellow) shows values from 0.3 to 0.5 and this means that they have a low positive correlation. Through this analysis, it is possible to check whether there is little relation of document similarity between image based and text based information retrieval. Besides, the OCR-text and the description have a negligible correlation. This means that the information extraction based on image processing could provide different results when compared to text based extraction. Also, even if the OCR-text and Num-text have a moderate correlation, their results are different when comparing with the Full-text. On average, the OCR-text shows a correlation value lower than the Num-text. Therefore, using the text extracted by OCR and the text related to images could be useful in patent information retrieval.

5. CONCLUSION

A patent image is one of the important elements in a patent document. Many patent experts use patent images for patent retrieval or analysing a patent. However, most often images need to be assessed manually because image processing technology is not readily available for patent retrieval and analysis systems. So, almost all research for patent retrieval is based on text and meta data elements based on text such as the IPC class.

This study suggests two image based methods for patent retrieval using image processing. One method uses the SIFT algorithm and the other method integrates OCR. The first method is to extract features from images by using the SIFT algorithm and then to calculate the similarity of documents through matching features. The other method is to extract text from images by using OCR and then calculate the similarity of patent documents through the existing text based method. Through comparing the document similarity between the suggested methods and a standard text based method, it is possible to check whether the new methods could be used for the patent retrieval and how different the result would be.

According to our results, the relationship between two methods and the existing methods are small. The correlation values are negligible. The method of using the SIFT-algorithm shows the largest difference to the standard text based method. The other method to use OCR also shows a difference with the existing method. Overall, the difference of the results between image based methods and text based method for patent retrieval is evaluated. The methods could prove useful for patent retrieval. They should be integrated into future patent search and analysis systems.

However, there are some drawbacks in the experiment. First, the accuracy of OCR is low for images in patents. Patent images are sometimes hand drawn. Besides, there is no standard format for the letters. This leads to many kinds of handwriting in patent images. If various handwriting styles in the patent image could be previously trained and used for the pattern matching, it would be possible to improve the quality of the OCR output. However, it is unlikely that such training can be provided.

Second, the SIFT-algorithm can be applied in different ways. In this study, the SIFT algorithm was used in a straightforward way. There is research using the SIFT-algorithm to classify different types of images. This led to high accuracy of over 95% [3]. If the document similarity based on the SIFT algorithm is calculated for each type of image after classification images, it could be possible to improve the result.

Currently, image processing is using more and more methods from Deep Learning which could also be applied to images in patents.

ACKNOWLEDGEMENTS

This paper was studied by supporting of PaiChai University.

REFERENCES

[1] WIPO, “What is intellectual property?,” World Intellectual Property Organization(WIPO), 2011. http://www.wipo.int/edocs/pubdocs/en/intproperty/450/wipo_pub_450.pdf (last retrieved : 01.0917)

[2] H. Joho, L. A. Azzopardi, and W. Vanderbauwhede, “A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements,” In Proceedings of the third symposium on information interaction in context, 2010, pp. 13-24.

[3] F. Piroi, M. Lupu, A. Hanbury, and V Zenz, “Clef-ip 2011: Retrieval in the intellectual property domain,” In The conference and labs of the evaluation forum - intellectual property(clef-ip), 2011.

[4] G. Csurka, J. M. Renders, and G. Jacquet, “Xrce’s participation at patent image classification and image-based patent retrieval tasks of the clef-ip 2011,” In The conference and labs of the evaluation forum – intellectual property(clefip), 2011.

[5] A. Hanbury, N. Bhatti, M. Lupu, and R. Mörzinger, “Patent image retrieval: a survey,” In Proceedings of the 4th workshop on patent information retrieval, 2011, pp. 3-8.

[6] P. Mandl, P. Womser-Hacker, Julia M. Struß, and Michael Schwantner, “Understanding trends in the patent domain,” In Proceedings of the first international workshop on patent mining and its applications (ipamin 2014) co-located with konvens 2014, 2014.

[7] A. Tiwari and V. Bansal, “Patseek: Content based image retrieval system for patent database,” In The fourth international conference on electronic business (iceb), 2004, pp. 1167-1171.

[8] C. A. Gonçalves, C. T. Gonçalves, R. Camacho, and E. C. Oliveira, “The impact of pre-processing on the classification of medline documents,” In Pattern recognition in information systems(pris), 2010, pp. 53-61.

[9] C. D. Manning, R. Raghavan, and H. Schütze, *Introduction to informationretrieval*, Cambridge University(Kyo-bo), 2008.

www.kci.go.kr

[10] P. C. Gaigole, L. Patil, and P. Chaudhari, "Preprocessing techniques in text categorization," International Journal of Computer Applications(IJCA), no. 3, 2013, pp. 1-3.

[11] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," Journal of documentation, vol. 60, no. 5, 2004, pp. 503-520.

[12] T. Maintz, "Digital and medical image processing. Universiteit Utrecht," 2005, pp. 247-272. http://www.cs.uu.nl/docs/vakken/ibv/reader/readerINFOIBV.pdf (last retrieved : 01.09.2017)

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, 2004, pp. 91-110.

[14] M. Brown and D. G. Lowe, "Invariant features from interest point groups," In Bmvc., 2002.

[15] C. Harris and M. Stephens, "A combined corner and edge detector," In Alvey vision conference, vol. 15, 1988, p. 50.

[16] A. Singh, K. Bacchuwar, and A. Bhasin, "A survey of ocr applications," International Journal of Machine Learning and Computing, vol. 2, no. 3, 2012, p. 314.

[17] R. Smith, "An overview of the tesseract ocr engine," Institute of Electrical and Electronics Engineers(IEEE), vol. 2, 2007, pp. 629-633.

[18] S. Vrochidis, "Papadopoulos, Symeon; Moumtzidou, Anastasia; Sidiropoulos, Panagiotis; Pianta, Emanuelle; Kompatsiaris, Ioannis (2010): Towards Content-based Patent Image Retrieval. A Framework Perspective," World Patent Information, vol. 32, no. 2, 2010, pp. 94-106.

[19] H. Ni, Z. Guo, and B. Huang, "Binary Patent Image Retrieval Using the Hierarchical Oriented Gradient Histogram," In: Proceedings of the International Conference on Service Science, ICSS, Weihai, China, May. 8-9 May, IEEE, 2015, pp. 23-27.

[20] D. Liparas, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris, "Concept-oriented Labelling of Patent Images Based on Random Forests and Proximity-driven Generation of Synthetic Data," In: Proceedings of the 25th International Conference on Computational Linguistics. COLIN. Dublin, Ireland, Aug. 23-29, 2014, pp. 25-32.

[21] S. Vrochidis, A. Moumtzidou, and I. Kompatsiaris, "Concept-based Patent Image Retrieval," In: World Patent Information, vol. 34, no. 4, 2012, pp. 292-303. DOI: 10.1016/j.wpi.2012.07.002

[22] N. Bhatti and A. Hanbury, "Image search in patents," A review In: IJDAR, vol. 16, no. 4, 2013, pp. 309-329. DOI: 0.1007/s10032-012-0197-5

[23] Arnaud Le Hors, Philippe Le Hégaret, Lauren Wood, Gavin Nicol, Jonathan Robie, Mike Champion, and Steve Byrne, "Document object model (dom) level 3 core specification," World Wide Web Consortium, 2004. http://travesia.mcu.es/portalnb/jspui/bitstream/10421/7473/1/DOM3-Core.pdf.(last retrieved : 01.09.2017)

[24] GridlogicsTechnologies, *Contact lenses technology insight report*, Patent iNSHIGHT Pro., 2014.

[25] E. J. Sarver and D. R. Sanders, "Aspheric, astigmatic, multifocal contact lens with asymmetric point spread function," Google Patents (US Patent App. 13/237,617), Feb. 9, 2012.

[26] WIPO, "Standard st. 36," World Intellectual Property Organization(WIPO), 2007. http://www.wipo.int/export/sites/www/standards/en/pdf/03-36-01.pdf (last retrieved : 01.0917)

[27] P. Chakraborty and A. Mallik, "An open source tesseract based tool for extracting text from images with application in braille translation for the visually impaired," International Journal of Computer Applications, vol. 68, no. 16, 2013.

[28] K. EL GAJOUI, F. A. ALLAH, and M. OUMSIS, "Training tesseract tool for amazigh ocr," In Recent researches in applied computer science: Proceedings of the 15th international conference on applied computer science 2015, pp. 20-22.

[29] D. F. Broderick, A. T. Foppe, J. Santilli, and R. C. Tucker, "Method and system for or dering customized cosmetic contact lenses," Google Patents (US Patent 6,746,120), Jun. 2004.

[30] A. Sozykin and T. Epanchintsev, "Mipr-a framework for distributed image processing using hadoop," In Application of information and communication technologies (aict), 2015 9th international conference on, 2015, pp. 35-39.

[31] H. Jonathon, S. Sina, and D. David, "The openimaj tutorial," 2012, http://openimaj.org/tutorial-pdf.pdf (last retrieved : 01.09.2017)

[32] G. Kondrak, "N-gram similarity and distance," In International symposium on string processing and information retrieval, 2005, pp. 115-126.

[33] E. Keogh and A. Mueen, "Curse of dimensionality," In Encyclopedia of machine learning, Springer, 2011, pp. 257-258.

[34] E. D'hondt, S. Verberne, C. Koster, and L. Boves, "Text representations for patent classification," Computational Linguistics, vol. 39, no. 3, 2013, pp. 755-775.

[35] C. Im, *Text classification for patents : Experiments with unigrams, bigrams and different weighting methods,* Unpublished master's thesis, Paichai University and Hildesheim University, 2016.

[36] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-ofspeech tagging with a cyclic dependency network," In Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology, vol. 1, 2003, pp. 173-180.

[37] Anjali Ganesh Jivani, "A comparative study of stemming algorithms," International Journal of Computer Technology and Applications(IJCTA), vol. 2, no. 6, 2011, pp. 1930-1938.

[38] C. Shalizi, "Stochastic processes (advanced probability ii)," Carnegie Mellon Univerity, 2006, pp. 189-196. http://www.stat.cmu.edu/_cshalizi/754/2006/notes/lecture-28.pdf (last retrieved : 01.09.2017)

[39] M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," Malawi Medical Journal, vol. 24, no. 3, 2012, pp. 69-71.

www.kci.go.kr

Jeong Beom Park

He received B.S. in Information and Communication Engineering from PaiChai University, Korea. He received M.A in Information Science from University of Hildesheim, Germany, and PaiChai University, Korea in 2017. He is currently doing his job at IWAZ in Korea. His main work is R&D for information retrieval, machine learning and deep learning.

Thomas Mandl

He is professor for Information Science at the University of Hildesheim in Germany where he is teaching within the program International Information Management. He studied Information Science and Computer Science at the University of Regensburg in Germany, the University of Koblenz and at the University of Illinois at Champaign/Urbana, USA. He first worked as a research assistant at the Social Science Information Centre in Bonn, Germany. He received both a doctorate degree in 2000 and a post doctorate degree (Habilitation) in 2006 from the University of Hildesheim. His research interests include information retrieval, human-computer interaction and internationalization of information technology.

Do Wan Kim

He received the B.A., M.A. and Ph.D. in Informatics from University of Regensburg, Germany. He was senior researcher in ETRI. Since 1997, he is professor at PaiChai University. His research interests include semantic web technologies, artificial intelligence, software quality evaluation and software ergonomics.

www.kci.go.kr