IMPROVED OCR QUALITY FOR SMART SCANNED DOCUMENT MANAGEMENT SYSTEM

Phan Viet Anh[1], *Nguyen Duy Tung Khanh*[1],
Tran Manh Dat[1], *Pham Van Dan*[1]

Abstract

The quality of the document images is a crucial factor for the performance of an Optical Character Recognition (OCR) model. Various issues from the input data hinder the recognition success such as heterogeneous layouts, skewness and proportional fonts. This paper investigated several algorithms for data pre-processing including image deskewing, table and document layout analysis to improve the accuracy of the OCR model and then built an end-to-end scanned document management system. We verified the algorithms using a well-known OCR software namely Tesseract. The experiments on a real dataset shown that our methods can accurately process document images with arbitrary angles of rotation, and different layouts. As a result, the accuracy by words of Tesseract can boost 23% for documents with complex structures. The quality of the output text allows to build a system to store and search documents efficiently.

Index terms

Optical Character Recognition (OCR); Table Recognition; Image Deskewing; Document Layout Analysis

1. Introduction

Optical character recognition (OCR) is converting images of documents of typed, handwritten or scanned text into machine-encoded text. OCR systems have been widely used in many practical applications such as invoice management [1], [2], CAPTCHA recognition [3], [4], building digital libraries [5], [6], and number plate recognition [7], [8]. The high quality of input data is one of the key factors to improve the recognition performance and thus affects the applicability of OCR systems.

Building an accurate OCR engine is a challenging problem. Many issues related to the input images that hinder OCR systems from achieving a high character recognition rate [9]. For example, noises, different font sizes and types, and skewing lead to errors in separating characters [10], [11]. Thus, the character-based algorithms can not work well. Moreover, heterogeneous layouts of documents containing tables, columns will degrade

[1]Le Quy Don Technical University

the performance of encoder-decoder based deep neural networks, e.g. Tesseract, that recognize the whole lines of text [12].

This paper aims to enhance the accuracy of OCR engines by pre-processing input data and build a searchable system for electronic documents. Our work focuses on processing several document types that are commonly appeared in the business work of government departments in Vietnam. Based on data observation, we have found that most of documents have high resolution and clean background so no contrast enhancement and background subtraction methods are needed. We developed the pre-processing techniques including image deskewing, table recognition, and document layout analysis. Applying the techniques will provide the input with quality sufficient enough for the OCR engine. The experiments on a real dataset of electronic documents shown that our pre-processing techniques can boost the accuracy of the OCR engine significantly. For application purpose, we built a system for storing, indexing and searching scanned documents to support the operation work of some agencies and organizations in Vietnam.

In summary, this paper makes the following contributions:

- Applying three pre-processing techniques to enhance the accuracy of OCR engines.
- Building an electronic document management system (eDMS) to promote the business work of companies and agencies.

The rest of this paper is organized as follows: Section 2 surveys studies related to pre-processing techniques for OCR engines. The proposed methods are described in Section 3. Section 4 presents the dataset, measurement and experimental results with discussions. Section 5 concludes our work and findings.

2. Related work

Converting document images to text has a wide range of real applications such as recognition and information extraction for business documents (passports, invoices, and bank statements) [13], [14]. Although various efforts to improve OCR performance, there is no universal solution for all electronic document types with different quality such as blurred, skewed, rotated, and complex structures [15]. In [16], Shen et al. tried to separate objects from the background. The purpose is to remove image background before feeding into the OCR engine. This helps to reduce noises in input images and hence improve the OCR performance. Following noise reduction approaches, Ye et al. proposed a method for text identification in images and video frames based on Support Vector Machines (SVMs) [17]. This method can process images with complex background to only extract text. Similarly, Shivananda et al. presented a hybrid model for separating text from the complex background [18]. The model combines connected components analysis and an unsupervised thresholding.

According to each kind of documents, many solutions have been investigated to obtain a high recognition rate. Brisinello et al. applied four different preprocessing methods to boost Tesseract's performance on images with low quality, low resolution and colorful background [19]. In [20], Bhagvati et al. introduced some important factors to help OCR system achieve high accuracy on Telugu and other Indian scripts. The factors were determined based on the characteristics of these characters.

For documents containing tables, Naganjaneyulu et al. proposed a heuristic-based table detection algorithm using hough lines and harris corner [21]. The main drawback of this algorithm is time-consuming. Shafait et al. used components of the layout analysis module of Tesseract to locate tables in documents [22]. This work only focuses on locating tables in document images, does not reconstruct the table structures in the output.

Recently, many researchers have applied deep learning-based methods for table detection and reconstruction. To locate tables, Gilani et al. used a region proposal network followed by a fully connected neural network [23]; Qasim et al. proposed a graph network [24]. Schreiber et al. [25] detected tables using Faster R-CNN [26] and semantic segmentation [27] for structure analysis. In [28], Paliwal et al. presented an end-to-end model for both table detection and structure analysis. The main drawback of deep learning-based methods is the need of a large amount of labeled data and computational time.

This work aims to process document images that may be rotated, skewed and contain tables. For rotated images, Hough transformation [29] is adopted to adjust the document orientation. For documents with tables, we need to perform two tasks including table detection and structure analysis. The details of our proposed methods will be presented in the next section.

3. Proposed methods

This section will describe our methods for input data normalization to improve accuracy of OCR engines. Generating accurate text is an important factor to leverage OCR model to practical applications. The data were collected from several companies and agencies. After observing, we have found that the scanned documents have different quality and can not feed directly to the recognition system. To address the main issues, we investigated the pre-processing techniques including image deskewing, table and layout analysis. Figure 1 shows the flow to combine such techniques to normalize the input images.

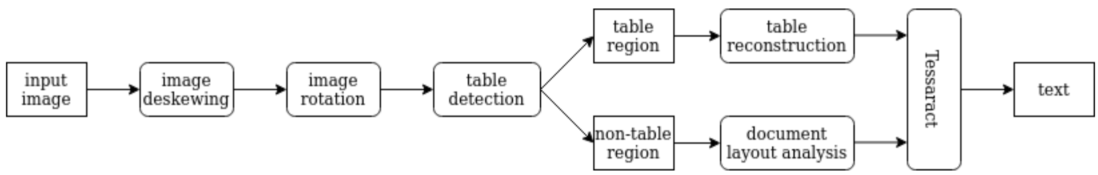
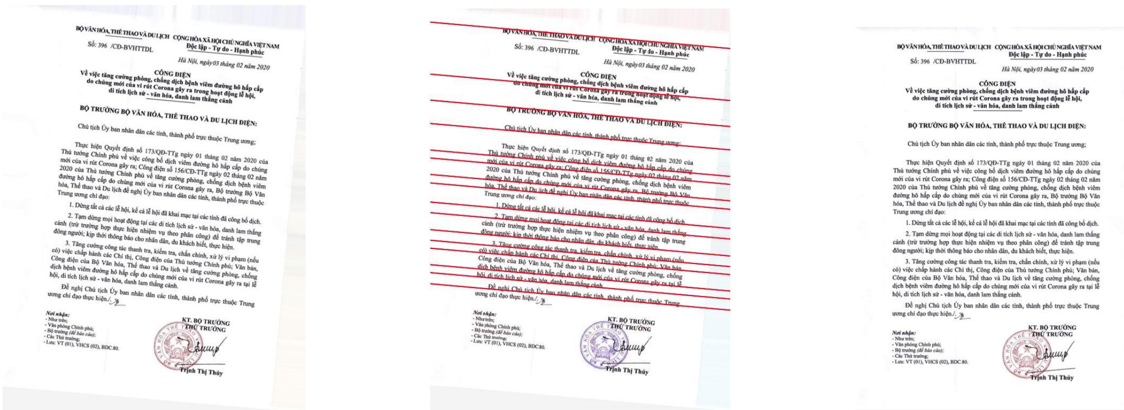


Fig. 1. The pipeline of the text recognition system with input data pre-processing.

3.1. Image deskewing

3.1.1. Image deskewing: Scanned documents usually skewed because they were not placed correctly on a flatbed scanner. This seriously affects the accuracy and speed of the OCR. Therefore, detecting and correcting the skew of scanned images are one of the crucial parts in OCR systems. This process is called image deskewing. To deskew scanned documents, we apply the Hough transform algorithm [29] to locate text lines in the images. This can be achieved by selecting appropriate parameters and filtering redundant lines. After that, we estimate the skew angle and make a rotation to align the document with four corners of the image. Figure 11 describes the entire scanned image deskewing process as mentioned above.



(a) (b) (c)

Fig. 2. Image deskewing process: (a) Input image, (b) Text lines detection using Hough transform, and (c) Output image

3.1.2. Page orientation correction: After rotating by the angle of text lines, the page orientation may be upright or upside down. The orientation now is estimated using algorithms in [30] and then we adjust the page to the correct position.

3.2. Table analysis

The purpose is to extract table components in the document to recognize separately and reconstruct in the output file. For encoder-decoder based OCR models that encode and generate the whole text lines instead of single characters, the scanned documents containing tables make a high error rate. The reason is that a line may contain text fragments of different cells and a cell may have some segments of text lines. This makes the decoder difficult to predict the output text and arrange the content. To address the issue, we extract sub images of each single cell to feed into the recognition model.

The steps to split a table into cells includes 1) locating the table, 2) finding cell vertices, and 3) determining the table structure. The rest of this section will describe more details about our method.

3.2.1. Table detection: To locate, we detect all lines in the images and then predict the set of lines that may form the table. Table lines are filtered by using image morphology operators [31] with appropriate structuring elements. This method is selected because tables are composed by vertical lines and horizontal lines. To apply the operators, we used dilation to highlight both vertical and horizontal lines in the image. Figure 3 illustrates using the dilation operator and a structuring element to emphasize vertical lines on an image. $B_{w,h}$ denotes the structuring element named B with the width and height of w and h respectively. In Figure 3, w is 1 and h is 3. The red point in B shows the origin of the structuring element. It can be seen that the dilation image has grown upwards and downwards compared to A. Additionally, the bigger h, the longer the vertical lines are. We use the structuring element with the width greater than the high for horizontal line detection and the width smaller than the high for vertical line detection. An example of table detection is shown in Figure 4.

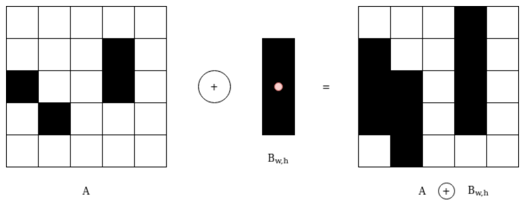


Fig. 3. Dilation of image A by structuring element B

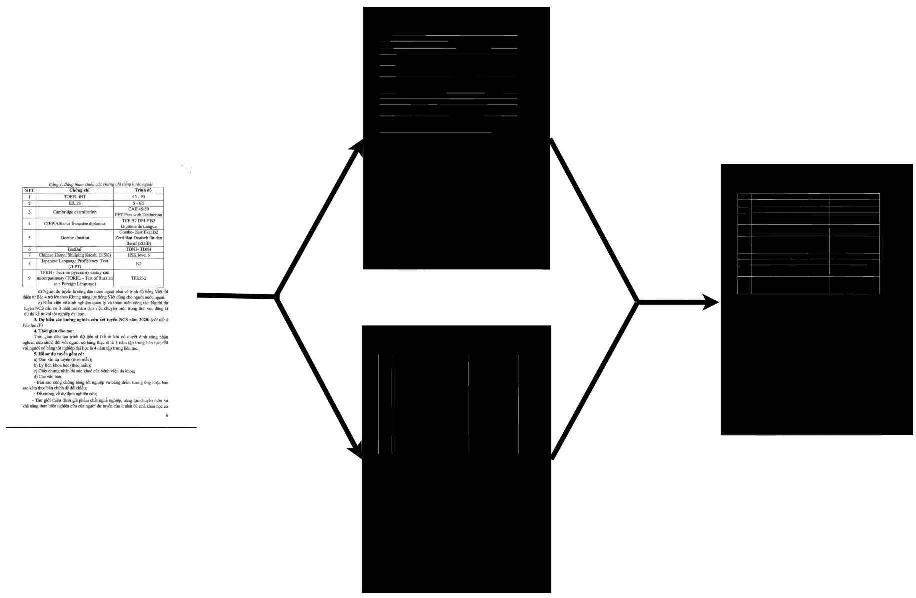


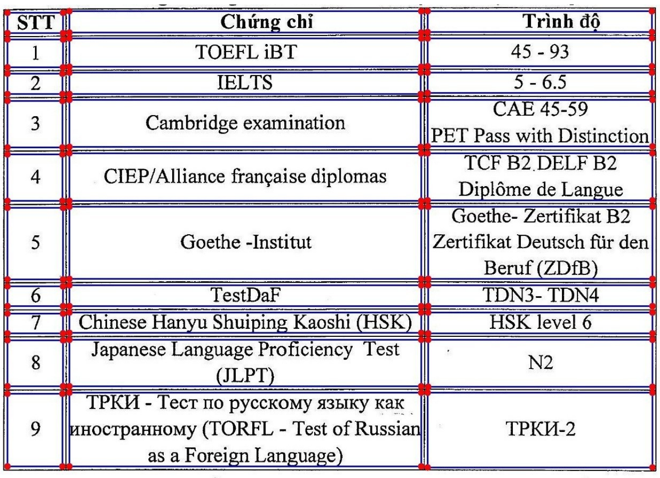
Fig. 4. Table detection using image morphology

3.2.2. Cell extraction and table construction: After locating tables, sub regions of cells are extracted. A table may have heterogeneous structures in which a cell may be the result of merging several cells. Thus we need to analyze the table structure to feed each single cell into the recognition engine, then output the texts into the similar format. The table analysis process has three main steps including 1) finding bounding rectangles, 2) merging cell corners, and 3) line alignment among cells.

Bounding rectangles. *Canny* algorithm [32] is applied to filter the edges in the table region. We find the inner contours and the consider the bounding rectangle of each contour as a cell. Figure 5 shows an example of cell extraction for a table. We denote S as the set of the rectangle vertices.

$$S = v_{i,j}, , i = \overline{1,n} \text{ and } j = \overline{1,4} \tag{1}$$

where n is the number of cells and $v_{i,j}$ denotes the j^{th} vertex of the i^{th} cell. These vertices are used to construct the table layout for the output text.



STT	Chứng chỉ	Trình độ
1	TOEFL iBT	45 - 93
2	IELTS	5 - 6.5
3	Cambridge examination	CAE 45-59 PET Pass with Distinction
4	CIEP/Alliance française diplomas	TCF B2 DELF B2 Diplôme de Langue
5	Goethe -Institut	Goethe- Zertifikat B2 Zertifikat Deutsch für den Beruf (ZDfB)
6	TestDaF	TDN3- TDN4
7	Chinese Hanyu Shuiping Kaoshi (HSK)	HSK level 6
8	Japanese Language Proficiency Test (JLPT)	N2
9	ТРКИ - Тест по русскому языку как иностранному (TORFL - Test of Russian as a Foreign Language)	ТРКИ-2

Fig. 5. Table cells detection

Merging cell corners. The vertex set S is reduced by merging points at each corner. Because cells are bounded by the inner rectangles (Figure 5), the corner points of adjacent cells are not identical. To merge such points, we first compute the Euclid distance among elements in S. Then, the vertices having the distances less than a threshold Δd are considered to belong to the same position. Δd is estimated according to the gap of text lines at image deskewing stage. Figure 6 illustrates the vertex merging process, where the vertices in the dashed circles are merged.

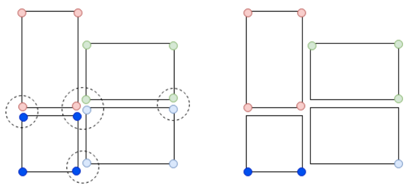


Fig. 6. Vertex merge process.

Line alignment among cells. After vertex merging, we determine all vertical and horizontal lines of the tables based on the cell vertex coordinates as in Figure 7.

Algorithm 1 presents the steps to construct the table from the vertex set. Starting points of vertical lines are called top anchor vertices and highlighted in red in Figure 7. Similarly, left anchor vertices are the starting points of horizontal lines, highlighted in green. The Algorithm 1 takes the vertex set as the input and find all top anchor and left anchor vertices. This process is illustrated in Figure 7. As described in Algorithm 1, the x coordinate of a vertex is ignored if its distance along x axis to any left anchor vertex is less than the threshold Δt_x. Similarly, we use the threshold Δt_y to remove non-top anchors. Δt_x and Δt_y are estimated from the gap of text lines and shared the same value. The algorithm starts from a top-left vertex, and collects all top and left anchors.

Algorithm 1 Table reconstruction

INPUT: Vertices set $V = v_1, v_2 \ldots v_N$, V_m is top-left vertex
OUTPUT: Top anchor vertices set V_x, left anchor vertices set V_y

Initialize anchor vertices: $V_x = \{V_m\}$, $V_y = \{V_m\}$
for V_i in V **do**
  for V_j in V_x **do**
    if $\|V_{jx} - V_{ix}\| < \Delta t_x$ **then**
      continue
    else
      $V_x = V_x \cup \{V_i\}$
    end if
  end for
  for V_k in V_y **do**
    if $\|V_{ky} - V_{iy}\| < \Delta t_y$ **then**
      continue
    else
      $V_y = V_y \cup \{V_i\}$
    end if
  end for
end for

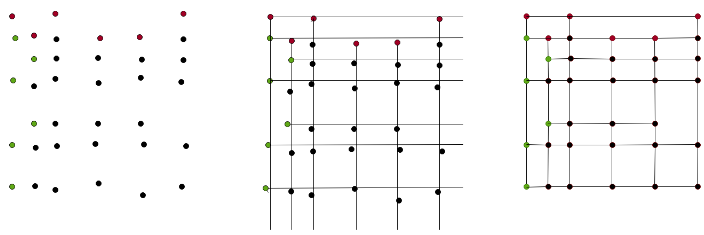
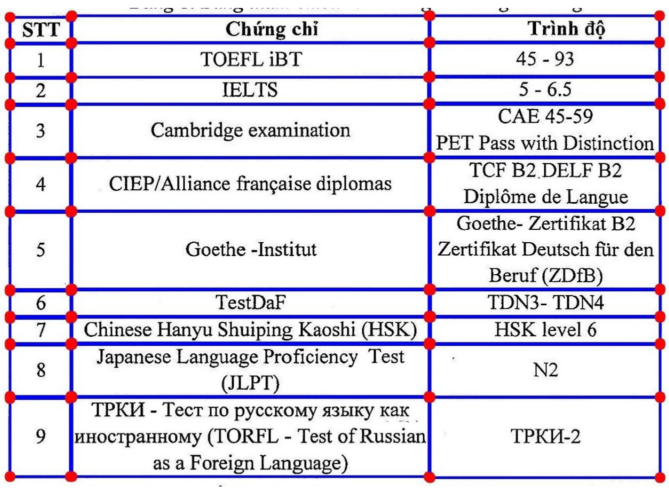


Fig. 7. Table reconstruction

Finally, the exact structure of the table is determined. We create a set of table lines by connecting all top anchor and left anchor vertices. Given any point in the vertex set,



STT	Chứng chỉ	Trình độ
1	TOEFL iBT	45 - 93
2	IELTS	5 - 6.5
3	Cambridge examination	CAE 45-59 PET Pass with Distinction
4	CIEP/Alliance française diplomas	TCF B2 DELF B2 Diplôme de Langue
5	Goethe -Institut	Goethe- Zertifikat B2 Zertifikat Deutsch für den Beruf (ZDfB)
6	TestDaF	TDN3- TDN4
7	Chinese Hanyu Shuiping Kaoshi (HSK)	HSK level 6
8	Japanese Language Proficiency Test (JLPT)	N2
9	ТРКИ - Тест по русскому языку как иностранному (TORFL - Test of Russian as a Foreign Language)	ТРКИ-2

Fig. 8. Table construction result

based on the distance to these lines, we can find the line that the vertex belongs to. After this step, the region of each single cell is identified. The sub image corresponding with this region is fed into OCR engine to recognize the text in the cell. Figure 8 shows the result of table analysis.

3.3. Document Layout Analysis

The purpose of this step is to separate a document into paragraphs and a paragraph into text lines. We use *X-Y Cut* algorithm [33] that applies on the projection of the number of black pixels (in the case of white paper backgrounds) on the X and Y axes to split the components in the image. An example of the projection is shown in Figure 9. The separation based on the projection is illustrated in Figure 10.

Fig. 9. The projection of image on X and Y axis

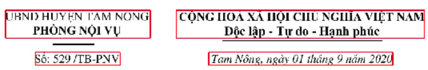


Fig. 10. Components separated based on the XY-cut algorithm

4. Experiments

4.1. Dataset

The dataset consists of 120 scanned images of Vietnamese documents dividing into two groups in which one contains tables (40 images) and one does not contain tables (80 images). Such two groups are called Table and Non-Table sets. We use the results on the documents containing tables to verify the quality of the table analysis method. The results on documents without tables are used to verify image deskewing, and layout analysis algorithms.

4.2. Evaluation Measures and Experimental Setting

To evaluate the performance of the methods, we use the measures of text similarity, and word error rate (WER) to estimate the distance between the ground truth and the predicted texts. To obtain the ground truth texts, we compared each scanned document and its OCR output to correct the errors.

The similarity of two texts is computed by difflib library[1]. Given two text T_1 and T_2, we find all matching blocks in which each block is defined as the form (i, j, n) such that $T_1[i : i + n] == T_2[j : j + n]$. The Similarity measure then is computed as follows:

$$Similarity = \frac{2 \times \sum_{i=1}^{K} |s_i|)}{|T_1| + |T_2|} \tag{2}$$

where K is the number of the matching blocks and $|s|$ denotes the length of the sequence s.

Our preprocessing methods are verified using Tesseract 4.0 that enables line recognition using LSTM networks. The experiments are to compare our preprocessing methods with those of Tesseract.

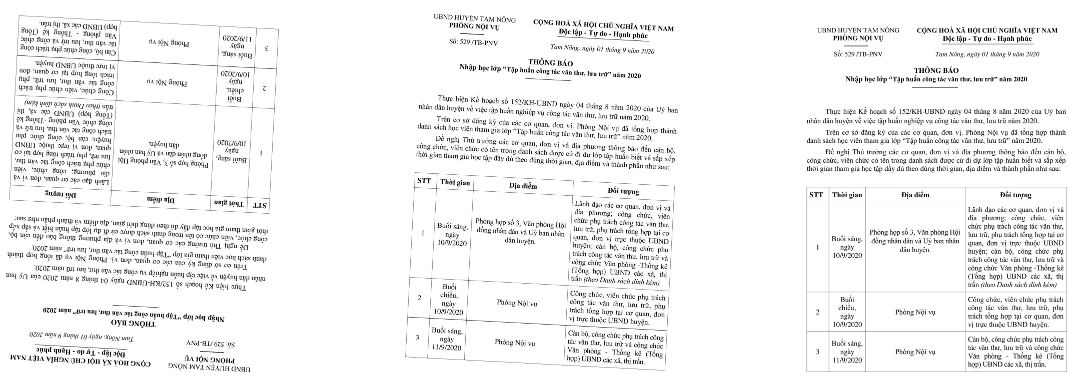
4.3. Results and Discussion

Table 1 compares the OCR accuracy according to the Similarity and WER in two cases with and without applying our proposed methods (eDMS) for Tesseract on Table dataset. For this dataset, we applied all the techniques including deskew, table and layout analysis. Our preprocessing methods improve Tesseract significantly. Specifically, the Similarity score is enhanced 0.23, and WER is reduced 23%. Figures 12 and 13 show the Similarity and WER for each document. As can be seen, our methods boost the accuracy of all the documents according to both Similarity and WER. Specially, several documents are unable to process by Tesseract resulting in very low performance, e.g. the third and twentieth documents. By applying our methods, Tesseract can recognize accurately.

[1] https://docs.python.org/3/library/difflib.html

To verify the deskew algorithm, we rotated the documents with different angles and try the recognition engine. Figures 14 and 15 shows Similarity and WER with different angles for two scenarios with and without application our deskew algorithm. It should be noted that our method can detect any rotated angle while Tesseract (which has also included image rotation as a preprocessing method) only works with angles around 0° and 270° (±4°).

We also compared our deskew algorithm with that in Tesseract. Figure 11 shows an example in which our method is more efficient.



(a) (b) (c)

Fig. 11. Image deskewing (a) Input image, (b) Deskew method in Tesseract, and (c) Proposed deskew method

To sum up, preprocessing data is essential for OCR engine to process non-standard input. This work present several techniques including deskew, table and layout analysis. These techniques are beneficial for Tesseract, a text line-based OCR recognition to process the several type of documents.

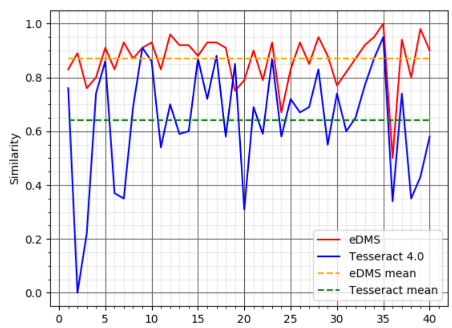


Fig. 12. Similarity on 40 images of Table dataset

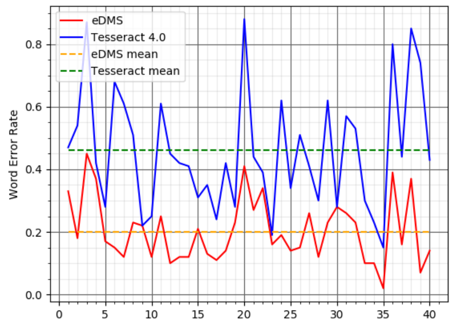


Fig. 13. WER on 40 images of Table dataset

Table 1. The performance of our proposed methods on the Table dataset

Measure	Tesseract 4.0	eDMS
similarity	0.64 ± 0.21	$\mathbf{0.87 \pm 0.09}$
WER	0.46 ± 0.19	$\mathbf{0.2 \pm 0.1}$

Error analysis. We observed the results and analyzed the characteristics of input images that our preprocessing methods are unable to correct. Figure 16 shows the failure cases including (a) containing seals, (b) mixing printed and handwritten characters, (c) containing noise lines causing by the scan process, and (d) blur table lines.

4.4. A smart scanned document management system

After obtaining the correct contents, we build a management system that enables to store and search scanned documents by text conveniently. This system is beneficial to the business work of various agencies and organizations where they archive a huge amount of paper documents. As an example, we surveyed an agencies and found that there are 15GB of scanned documents in recent two years.

The architecture of the system is shown in Figure 17. Given a paper document, after scanning and uploading, the system will convert the image to the text. A document then is stored in a tube of the scanned image and the OCR text. The system allows users texting to search and return both the original image and the content. To search efficiently, we use Elasticsearch[2], a highly scalable open-source full-text search and analytic engine.

5. Conclusion

This paper presented three image preprocessing methods to improve the OCR performance for scanned documents. The experimental results have shown that our methods

[2]https://www.elastic.co/

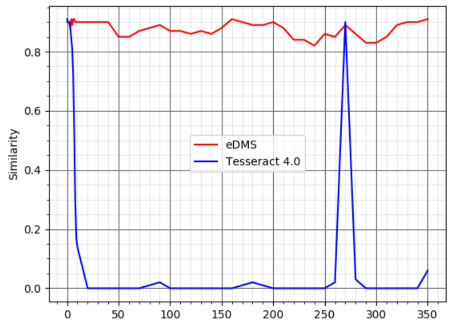


Fig. 14. Similarity on non-Table dataset with different skew angle

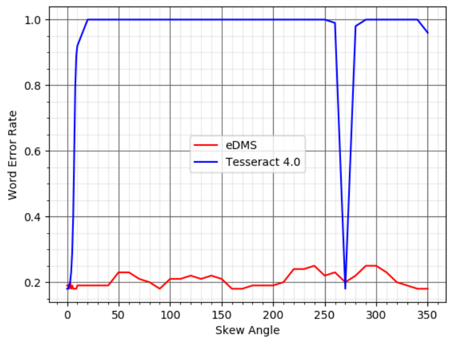


Fig. 15. WER on non-Table dataset with different skew angle

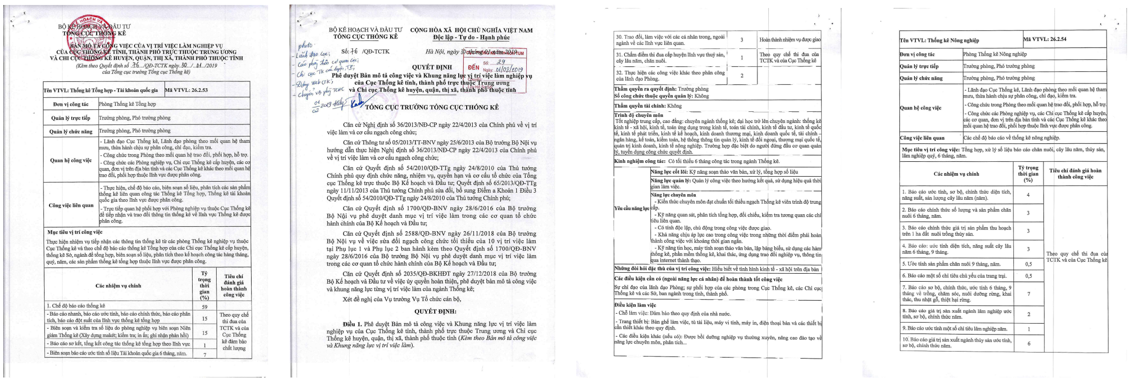
can process documents rotated by arbitrary angles and analyze tables with complex structures. As a result, the method boosts Tesseract significantly. The paper also introduced a smart scanned document management system that supports the paper work of many agencies and organizations.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2018.306.

References

[1] K. Kohlmaier, E. Hess, and B. Klehr, "Invoice verification process," Apr. 27 2006, uS Patent App. 11/026,026.
[2] H. T. Ha, Z. Nevěřilová, A. Horák *et al.*, "Recognition of ocr invoice metadata block types," in *International Conference on Text, Speech, and Dialogue*. Springer, 2018, pp. 304–312.
[3] P. Lupkowski and M. Urbanski, "Semcaptcha—user-friendly alternative for ocr-based captcha systems," in *2008 International Multiconference on Computer Science and Information Technology*. IEEE, 2008, pp. 325–329.



(a) (b) (c) (d)

Fig. 16. Some failure cases (a) containing a seal (b) mixing printed and handwritten characters, (c) noise lines (d) blur lines.

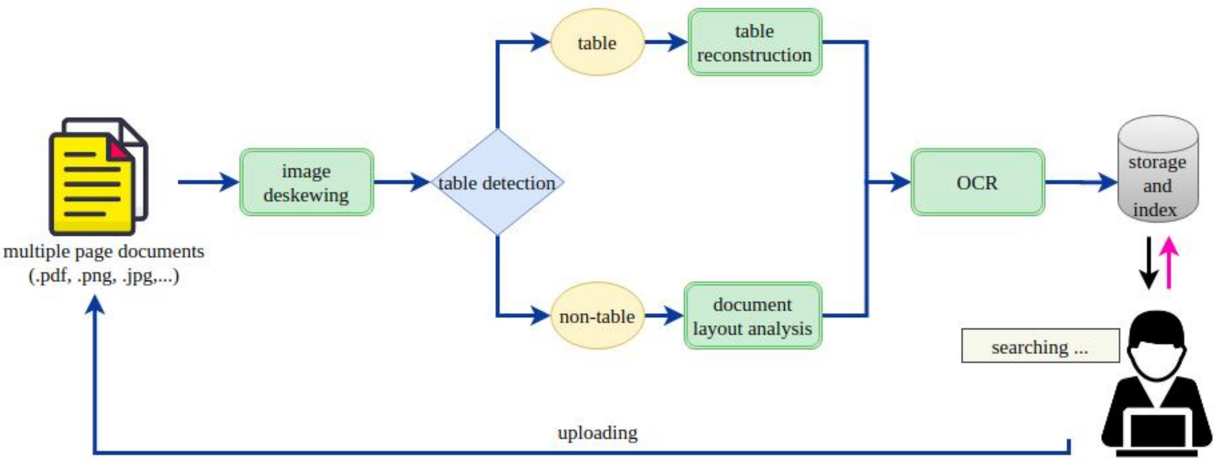


Fig. 17. The eDMS architecture

[4] D. Lin, F. Lin, Y. Lv, F. Cai, and D. Cao, "Chinese character captcha recognition and performance estimation via deep neural network," *Neurocomputing*, vol. 288, pp. 11–19, 2018.

[5] G. Chiron, A. Doucet, M. Coustaty, M. Visani, and J.-P. Moreux, "Impact of ocr errors on the use of digital libraries: towards a better access to information," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2017, pp. 1–4.

[6] L. Zhang and C. L. Tan, "Warped image restoration with applications to digital libraries," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 2005, pp. 192–196.

[7] E. K. Kaur and V. K. Banga, "Number plate recognition using ocr technique," *International Journal of Research in Engineering and Technology*, vol. 2, no. 09, p. 286290, 2013.

[8] M. T. Qadri and M. Asif, "Automatic number plate recognition system for vehicle identification using optical character recognition," in *2009 International Conference on Education Technology and Computer*. IEEE, 2009, pp. 335–338.

[9] A. Gupta, R. Gutierrez-Osuna, M. Christy, B. Capitanu, L. Auvil, L. Grumbach, R. Furuta, and L. Mandell, "Automatic assessment of ocr quality in historical documents," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Citeseer, 2015.

[10] R. Holley, "How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs," *D-Lib Magazine*, vol. 15, no. 3/4, 2009.

[11] K. A. Hamad and M. Kaya, "A detailed analysis of optical character recognition technology," *International Journal of Applied Mathematics, Electronics and Computers*, vol. 4, no. 1, pp. 244–249, 2016.

[12] M. Diem, F. Kleber, and R. Sablatnig, "Text line detection for heterogeneous documents," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 743–747.

[13] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. Jawahar, "Icdar2019 competition on scanned receipt ocr and information extraction," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1516–1520.

[14] Y. Ishitani, "Model-based information extraction method tolerant of ocr errors for document images," *International Journal of Computer Processing of Oriental Languages*, vol. 15, no. 02, pp. 165–186, 2002.

[15] S. Kompalli, S. Nayak, S. Setlur, and V. Govindaraju, "Challenges in ocr of devanagari documents," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 2005, pp. 327–331.

[16] M. Shen and H. Lei, "Improving ocr performance with background image elimination," in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, 2015, pp. 1566–1570.

[17] Q. Ye, W. Gao, and Q. Huang, "Automatic text segmentation from complex background," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, vol. 5. IEEE, 2004, pp. 2905–2908.

[18] N. Shivananda and P. Nagabhushan, "Separation of foreground text from complex background in color document images," in *2009 Seventh International Conference on Advances in Pattern Recognition*. IEEE, 2009, pp. 306–309.

[19] M. Brisinello, R. Grbić, M. Pul, and T. Anđelić, "Improving optical character recognition performance for low quality images," in *2017 International Symposium ELMAR*. IEEE, 2017, pp. 167–171.

[20] C. Bhagvati, T. Ravi, S. M. Kumar, and A. Negi, "On developing high accuracy ocr systems for telugu and other indian scripts," in *Language Engineering Conference, 2002. Proceedings*. IEEE, 2002, pp. 18–23.

[21] G. Naganjaneyulu, N. V. Sathwik, and A. Narasimhadhan, "A multi clue heuristic based algorithm for table detection," in *2016 IEEE Region 10 Conference (TENCON)*. IEEE, 2016, pp. 1246–1249.

[22] F. Shafait and R. Smith, "Table detection in heterogeneous documents," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, 2010, pp. 65–72.

[23] A. Gilani, S. R. Qasim, I. Malik, and F. Shafait, "Table detection using deep learning," in *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 771–776.

[24] S. R. Qasim, H. Mahmood, and F. Shafait, "Rethinking table recognition using graph neural networks," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 142–147.

[25] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, "Deepdesrt: Deep learning for detection and structure recognition of tables in document images," in *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 1162–1167.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[28] S. S. Paliwal, D. Vishwanath, R. Rahul, M. Sharma, and L. Vig, "Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 128–133.

[29] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[30] R. Unnikrishnan and R. Smith, "Combined script and page orientation estimation using the tesseract ocr engine," in *Proceedings of the international workshop on multilingual OCR*, 2009, pp. 1–7.

[31] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE transactions on pattern analysis and machine intelligence*, no. 4, pp. 532–550, 1987.

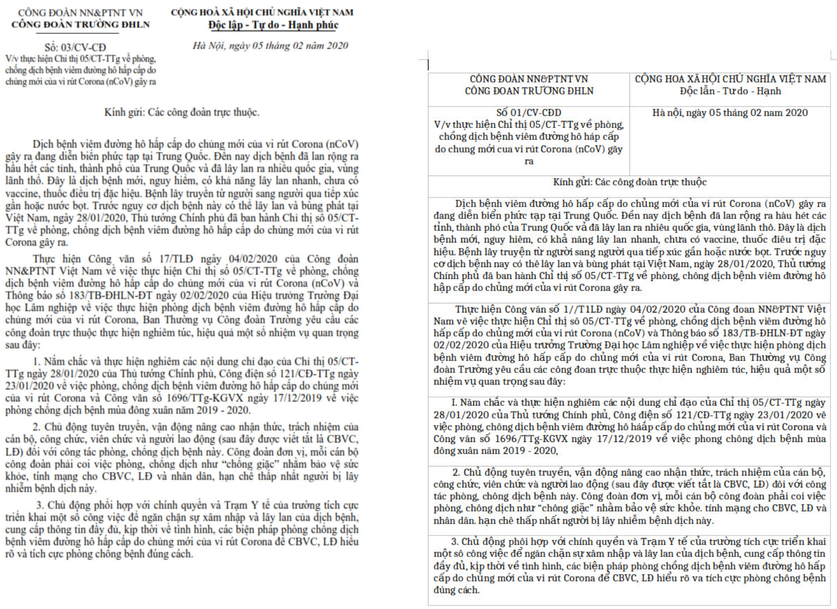
[32] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[33] F. Shafait and T. M. Breuel, "The effect of border noise on the performance of projection-based page segmentation methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 846–851, 2010.

Manuscript received 20-2-2020; Accepted 14-5-2020.

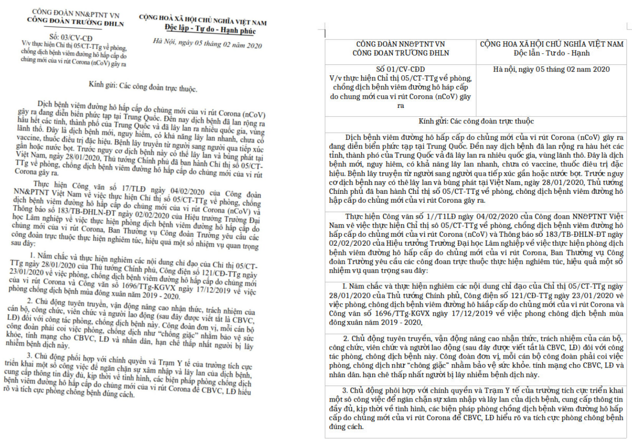
Appendix A
Results obtained from eDMS system

The output of the eDMS system for some scanned documents types are shown in Figures 18, 19, 20, and 21.



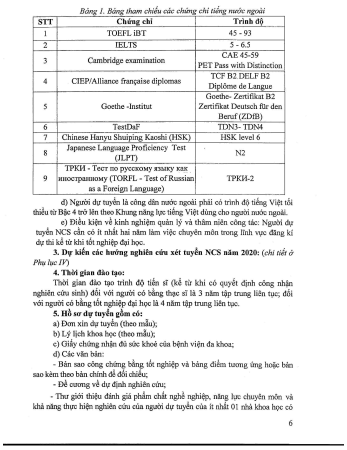
(a) (b)

Fig. 18. A document just contains text



(a) (b)

Fig. 19. A skewed document image



Bảng 1. Bảng tham chiếu các chứng chỉ tiếng nước ngoài

STT	Chứng chỉ	Trình độ
1	TOEFL iBT	45 - 93
2	IELTS	5 - 6.5
3	Cambridge examination	CAE 45-59 PET Pass with Distinction
4	CIEP/Alliance française diplomas	TCF B2 DELF B2 Diplôme de Langue
5	Goethe -Institut	Goethe- Zertifikat B2 Zertifikat Deutsch für den Beruf (ZDfB)
6	TestDaF	TDN3- TDN4
7	Chinese Hanyu Shuiping Kaoshi (HSK)	HSK level 6
8	Japanese Language Proficiency Test (JLPT)	N2
9	ТРКИ - Тест по русскому языку как иностранному (TORFL - Test of Russian as a Foreign Language)	ТРКИ-2

d) Người dự tuyển là công dân nước ngoài phải có trình độ tiếng Việt tối thiểu từ Bậc 4 trở lên theo Khung năng lực tiếng Việt dùng cho người nước ngoài.

e) Điều kiện về kinh nghiệm quản lý và thâm niên công tác: Người dự tuyển NCS cần có ít nhất hai năm làm việc chuyên môn trong lĩnh vực đăng kí dự thi kể từ khi tốt nghiệp đại học.

3. Dự kiến các hướng nghiên cứu xét tuyển NCS năm 2020: (*chi tiết ở Phụ lục IV*)

4. Thời gian đào tạo:

Thời gian đào tạo trình độ tiến sĩ (kể từ khi có quyết định công nhận nghiên cứu sinh) đối với người có bằng thạc sĩ là 3 năm tập trung liên tục; đối với người có bằng tốt nghiệp đại học là 4 năm tập trung liên tục.

5. Hồ sơ dự tuyển gồm có:

a) Đơn xin dự tuyển (theo mẫu);

b) Lý lịch khoa học (theo mẫu);

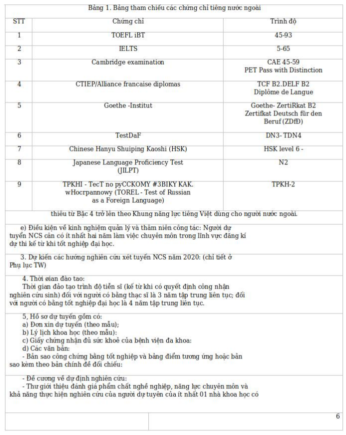
c) Giấy chứng nhận đủ sức khoẻ của bệnh viện đa khoa;

d) Các văn bản:

- Bản sao công chứng bằng tốt nghiệp và bảng điểm tương ứng hoặc bản sao kèm theo bản chính để đối chiếu;
- Đề cương về dự định nghiên cứu;
- Thư giới thiệu đánh giá phẩm chất nghề nghiệp, năng lực chuyên môn và khả năng thực hiện nghiên cứu của người dự tuyển của ít nhất 01 nhà khoa học có

6

(a)



Bảng 1. Bảng tham chiếu các chứng chỉ tiếng nước ngoài

STT	Chứng chỉ	Trình độ
1	TOEFL iBT	45-93
2	IELTS	5-65
3	Cambridge examination	CAE 45-59 PET Pass with Distinction
4	CTIEP/Alliance francaise diplomas	TCF B2.DELF B2 Diplôme de Langue
5	Goethe -Institut	Goethe- ZertiRkat B2 Zertifkat Deutsch für den Beruf (ZDfĐ)
6	TestDaF	DN3- TDN4
7	Chinese Hanyu Shuiping Kaoshi (HSK)	HSK level 6 -
8	Japanese Language Proficiency Test (JILPT)	N2
9	TPKHI - TecT no pyCCKOMY #3BIKY KAK. wHocrpannowy (TOREL - Test of Russian as a Foreign Language)	TPKH-2

thiểu từ Bậc 4 trở lên theo Khung năng lực tiếng Việt dùng cho người nước ngoài.

e) Điều kiện về kinh nghiệm quản lý và thâm niên công tác: Người dự tuyển NCS cần có ít nhất hai năm làm việc chuyên môn trong lĩnh vực đăng kí dự thi kế từ khi tốt nghiệp đại học.

3. Dự kiến các hướng nghiên cứu xét tuyển NCS năm 2020: (chỉ tiết ở Phụ lục TW)

4. Thời øian đào tao:
Thời gian đảo tạo trình độ tiễn sĩ (kế từ khi có quyết định công nhận nghiên cứu sinh) đối với người có bằng thạc sĩ là 3 năm tập trung liên tục; đối với người có bằng tốt nghiệp đại học là 4 năm tập trung liên tục.

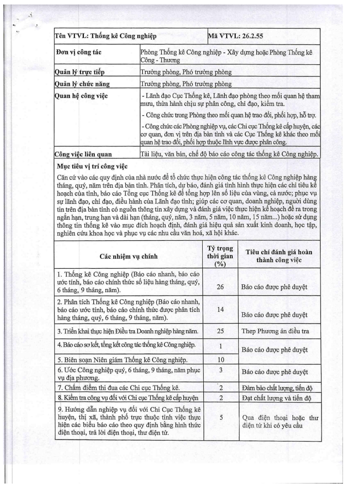
5, Hồ sơ dự tuyển gồm có:
a) Đơn xin dự tuyển (theo mẫu);
b) Lý lịch khoa học (theo mẫu):
c) Giấy chứng nhận đủ sức khoẻ của bệnh viện đa khoa:
d) Các văn bản:
- Bản sao công chứng bằng tốt nghiệp và bảng điểm tương ứng hoặc bản sao kèm theo bản chính để đối chiếu:

- Đề cương về dự định nghiên cứu:
- Thư giới thiệu đánh giá phẩm chất nghề nghiệp, năng lực chuyên môn và khả năng thực hiện nghiên cứu của người dự tuyển của ít nhất 01 nhà khoa học có

6

(b)

Fig. 20. A document contains tables.



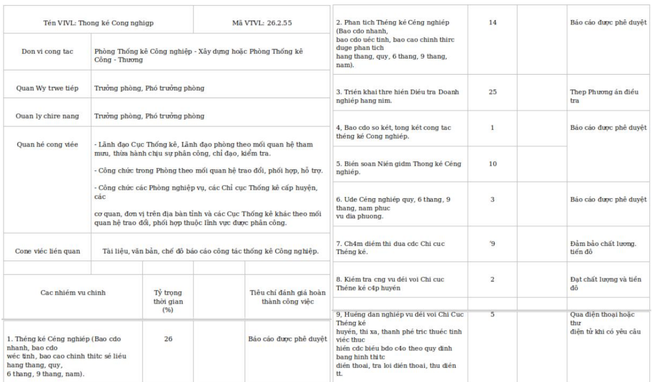
Tên VTVL: Thống kê Công nghiệp	**Mã VTVL: 26.2.55**
Đơn vị công tác	Phòng Thống kê Công nghiệp - Xây dựng hoặc Phòng Thống kê Công - Thương
Quản lý trực tiếp	Trưởng phòng, Phó trưởng phòng
Quản lý chức năng	Trưởng phòng, Phó trưởng phòng
Quan hệ công việc	- Lãnh đạo Cục Thống kê, Lãnh đạo phòng theo mối quan hệ tham mưu, thừa hành chịu sự phân công, chỉ đạo, kiểm tra. - Công chức trong Phòng theo mối quan hệ trao đổi, phối hợp, hỗ trợ. - Công chức các Phòng nghiệp vụ, các Chi cục Thống kê cấp huyện, các cơ quan, đơn vị trên địa bàn tỉnh và các Cục Thống kê khác theo mối quan hệ trao đổi, phối hợp thuộc lĩnh vực được phân công.
Công việc liên quan	Tài liệu, văn bản, chế độ báo cáo công tác thống kê Công nghiệp.

Mục tiêu vị trí công việc

Căn cứ vào các quy định của nhà nước để tổ chức thực hiện công tác thống kê Công nghiệp hàng tháng, quý, năm trên địa bàn tỉnh. Phân tích, dự báo, đánh giá tình hình thực hiện các chỉ tiêu kế hoạch của tỉnh, báo cáo Tổng cục Thống kê để tổng hợp lên số liệu của vùng, cả nước; phục vụ sự lãnh đạo, chỉ đạo, điều hành của Lãnh đạo tỉnh; giúp các cơ quan, doanh nghiệp, người dùng tin trên địa bàn tỉnh có nguồn thông tin xây dựng và đánh giá việc thực hiện kế hoạch đề ra trong ngắn hạn, trung hạn và dài hạn (tháng, quý, năm, 3 năm, 5 năm, 10 năm, 15 năm...) hoặc sử dụng thông tin thống kê vào mục đích hoạch định, đánh giá hiệu quả sản xuất kinh doanh, học tập, nghiên cứu khoa học và phục vụ các nhu cầu văn hoá, xã hội khác.

Các nhiệm vụ chính	Tỷ trọng thời gian (%)	Tiêu chí đánh giá hoàn thành công việc
1. Thống kê Công nghiệp (Báo cáo nhanh, báo cáo ước tính, báo cáo chính thức số liệu hàng tháng, quý, 6 tháng, 9 tháng, năm).	26	Báo cáo được phê duyệt
2. Phân tích Thống kê Công nghiệp (Báo cáo nhanh, báo cáo ước tính, báo cáo chính thức được phân tích hàng tháng, quý, 6 tháng, 9 tháng, năm).	14	Báo cáo được phê duyệt
3. Triển khai thực hiện Điều tra Doanh nghiệp hàng năm.	25	Thep Phương án điều tra
4. Báo cáo sơ kết, tổng kết công tác thống kê Công nghiệp.	1	Báo cáo được phê duyệt
5. Biên soạn Niên giám Thống kê Công nghiệp.	10	
6. Ước Công nghiệp quý, 6 tháng, 9 tháng, năm phục vụ địa phương.	3	Báo cáo được phê duyệt
7. Chấm điểm thi đua các Chi cục Thống kê.	2	Đảm bảo chất lượng, tiến độ
8. Kiểm tra công vụ đối với Chi cục Thống kê cấp huyện	2	Đạt chất lượng và tiến độ
9. Hướng dẫn nghiệp vụ đối với Chi Cục Thống kê huyện, thị xã, thành phố trực thuộc tỉnh việc thực hiện các biểu báo cáo theo quy định bằng hình thức điện thoại, trả lời điện thoại, thư điện tử.	5	Qua điện thoại hoặc thư điện tử khi có yêu cầu

(a)



Tén VIVL: Thong ké Cong nghigp	Mã VTVL: 26.2.55
Don vi cong tac	Phòng Thống kê Công nghiệp - Xây dựng hoặc Phòng Thống kê Công - Thương
Quan Wy trwe tiép	Trưởng phòng, Phó trưởng phòng
Ouan ly chire nang	Trưởng phòng, Phó trưởng phòng
Quan hé cong viée	- Lãnh đạo Cục Thống kê, Lãnh đạo phòng theo mối quan hệ tham mưu, thừa hành chịu sự phân công, chỉ đạo, kiểm tra. - Công chức trong Phòng theo mối quan hệ trao đổi, phối hợp, hỗ trợ. - Công chức các Phòng nghiệp vụ, các Chỉ cục Thống kê cấp huyện, các cơ quan, đơn vị trên địa bàn tỉnh và các Cục Thống kê khác theo mối quan hệ trao đổi, phối hợp thuộc lĩnh vực được phân công.
Cone viéc lién quan	Tài liệu, văn bản, chế đô báo cáo công tác thống kê Công nghiệp.

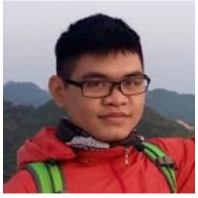
Cac nhiém vu chinh	Tỷ trọng thời gian (%)		Tiêu chí đánh giá hoàn thành công việc
1. Théng ké Céng nghiép (Bao cdo nhanh, bao cdo wéc tinh, bao cao chinh thitc sé liéu hang thang, quy, 6 thang, 9 thang, nam).	26		Báo cáo được phê duyệt

2. Phan tich Théng ké Céng nghiép (Bao cdo nhanh, bao cdo uéc tinh, bao cao chinh thirc duge phan tich hang thang, quy, 6 thang, 9 thang, nam).	14		Báo cáo được phê duyệt
3. Trién khai thre hién Diéu tra Doanh nghiép hang nim.	25		Thep Phương án điều tra
4, Bao cdo so két, tong két cong tac théng ké Cong nghiép.	1		Báo cáo được phê duyệt
5. Bién soan Nién gidm Thong ké Céng nghiép.	10		
6. Ude Céng nghiép quy, 6 thang, 9 thang, nam phuc vu dia phuong.	3		Báo cáo được phê duyệt
7. Ch4m diém thi dua cdc Chi cuc Théng ké.	'9		Đảm bảo chất lương. tiến đô
8. Kiém tra cng vu déi voi Chi cuc Théne ké c4p huyén	2		Đạt chất lượng và tiến đô
9, Huéng dan nghiép vu déi voi Chi Cuc Théng ké huyén, thi xa, thanh phé tric thuéc tinh viéc thuc hién cdc biéu bdo c4o theo quy dinh bang hinh thitc dién thoai, tra loi dién thoai, thu dién tt.	5		Qua điện thoại hoặc thư điện tử khi có yêu câu

(b)

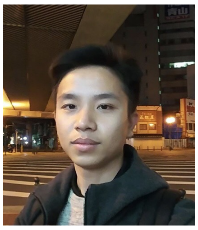
Fig. 21. A document with a complex structure table

Phan Viet Anh Dr. Phan Viet Anh received the B.Sc. degree in information technology, and the MSc degree in computer science from Le Quy Don Technical University, Vietnam, in 2008 and 2013, respectively, PhD degree in computer science from Japan Advanced Institute of Science and Technology (JAIST) in 2018 with outstanding student award. His research interests include machine learning, software engineering, evolutionary computation, and deep learning.



Nguyen Duy Tung Khanh graduated with a Bachelor of Engineering degree from Le Quy Don Technical University, majoring in Information Technology in 2018. He is currently teaching assistant at Department of Information Security, Faculty of Information Technology, Le Quy Don Technical University, Vietnam. His research interests include machine learning and information security.

Tran Manh Dat He is currently a student at Faculty of Information Technology, Le Quy Don Technical University, Vietnam. His current interests include computer vision, NLP and embedded systems for autonomous machines.



Pham Van Dan He is currently a student at Faculty of Information Technology, Le Quy Don Technical University, Vietnam. His current interests include computer vision and speech processing.

NÂNG CAO CHẤT LƯỢNG NHẬN DẠNG KÝ TỰ QUANG HỌC CHO HỆ THỐNG QUẢN LÝ VĂN BẢN THÔNG MINH

Tóm tắt

Chất lượng ảnh là yếu tố quan trọng đối với hiệu năng của mô hình Nhận dạng ký tự quang học (OCR). Các vấn đề khác nhau từ dữ liệu đầu vào cản trở sự thành công trong việc nhận dạng như bố cục không đồng nhất, độ lệch (ảnh bị xoay hoặc méo) và cỡ chữ khác nhau. Bài báo này đã nghiên cứu một số thuật toán tiền xử lý dữ liệu bao gồm khử lệch, phân tích cấu trúc bảng và bố cục tài liệu để nâng cao độ chính xác của mô hình OCR và sau đó xây dựng một hệ thống tổng thể cho việc quản lý tài liệu. Chúng tôi đã kiểm định các thuật toán bằng phần mềm OCR nổi tiếng là Tesseract. Các thử nghiệm trên tập dữ liệu thực cho thấy rằng các phương pháp của chúng tôi có thể xử lý chính xác hình ảnh tài liệu với các góc quay tùy ý và các bố cục khác nhau. Do đó, độ chính xác theo từ trong Tesseract có thể tăng 23 % đối với các tài liệu có cấu trúc phức tạp. Chất lượng của văn bản đầu ra cho phép xây dựng hệ thống lưu trữ và tìm kiếm văn bản một cách hiệu quả.