Finding Inconsistencies between Scanned Copies of Business Documents

O. A. Slavin
Federal Research Center "Computer Sciences and Control"
Russian Academy of Sciences
Moscow, Russia
oslavin@isa.ru

E. I. Andreeva [1,2]
[1]OOO "Smart Engines Service"
[2] "Moscow Institute of Physics and Technology
(State University), MIPT"
Moscow, Russia
andreeva@phystech.edu

***Abstract*— The paper considers methods for solving the problem of comparing digitized copies of business documents. Such a task arises when comparing two copies of documents signed by two parties in order to find possible modifications made by one party, for example, in the banking sector when concluding contracts in paper form. A method of comparison of two digitized images based on a combination of comparison methods when using text recognition algorithms and methods of comparison of segmented parts of the image using feature points is proposed. The proposed method of classification can be used in modern CAD to analyze the content of text documents.**

Keywords— comparison of documents; image segmentation; feature points; automatic text recognition; Levenshtein distance

I. INTRODUCTION

When concluding contracts in paper form, a situation may arise when one of the parties may make changes to its copy of the contract, which are undesirable for the opposite party, for example, in case of a change in the terms of the contract by the client of a bank in its favor. To avoid this, you need a document comparison procedure, the purpose of the comparison is to find modifications in the document.

Scanned pages consisting of lines, words and individual symbols were considered as objects for comparing images of business documents.

We have considered the following types of possible modifications:

- replace one or more characters in a word;
- replace one word with another;
- adding a word or a group of words;
- delete a word or a group of words.

The paper describes a method that allows to detect these modifications in scanned images of business documents.

II. METHOD

Let us consider the problem of finding inconsistencies between scanned images of two instances of business document pages. One image corresponds to the model page, the other one – to the test page. The input data are the model image with a markup on the words. In the test image, you the modifications must be selected.

Let us define a document as a set of lines $D = \left\{ L_i^d \right\}_{i=1}^{|D|}$, a line as a set of words $L^d = \left\{ W_j^l \right\}_{j=1}^{|L^d|}$, and a word as a rectangle on the raster $W^l =< x, y, w, h >$, where (x, y) are the coordinates of the upper left point of this rectangle, w is its width, and h is its height, W^l is the word on the line l, L^d is a string in the document d.

Two documents are considered to be free from modifications if all the lines in them are coordinated, the coordination of the lines is carried out by checking the coordination of the words that make up the lines. The definition of coordinated words is given below.

The nonconformance search method consists of the following steps:

- test image processing:
  - image pre-processing,
  - text-to-line segmentation,
  - segmentation of text into words,
- comparison of strings of model and test images, establishing matches between words and strings.

Two strings are considered coordinated if the proportion of coordinated words in these strings is higher than the specified threshold $line\_simil$.

An important step in the pre-processing of scanned images was the normalization of the image (turning), after which the text lines are in absolutely horizontal position. We used for it the fast transform Hough algorithm [3].

Next, the segmentation is carried out by building a histogram, where there are frames of words in the image. First, the connectivity component is selected, the main components are selected, horizontal and vertical histograms are formed, taking into account possible emissions. Vertical segmentation into lines and horizontal segmentation of the obtained lines into words is carried out by analyzing histograms and

978-1-5386-5824-6/18/$31.00 ©2018 IEEE

morphological operations [10]. After the segmentation, the images were presented as a set of coordinates of the word frames.

A combination of several methods is used to compare words.

The first method is based on the recognition of the text obtained by applying appropriate algorithms to the scanned images. We used two means of recognition:

- free software OCR Tesseract;
- recognition library used in software products, e.g. Smart ID Reader [6].

The work [2] represents lists the advantages of OCR Tesseract: the possibility of free distribution and presentation of recognition results in the format of HOCR (HTML OCR), which contains information about the coordinates of the frames of recognized words. Ordinary errors of the Tesseract OCR (we used version 4.0), such as wrong recognition of the structure of the page and the error recognition of characters' trouble to compare documents and make false positives in the results of the comparison.

The Smart Engines recognition library [7], the basic principles of which are described in the work [8], has less functionality than OCR Tesseract, provides more accurate recognition of documents, primarily Russian-speaking. Recognition in this library is based on two types of neural networks. The first neural network consists of several convolutional layers and two fully connected layers. Each convolutional layer is followed by a subsampling layer, and each subsampling layer and a fully connected layer is followed by a heuristic layer of random partial zeroing. Hyperbolic tangent was used as activation functions. In order to increase the accuracy of segmentation and character recognition was used a combination of convolutional network and two-way recurrent neural network. The recurrent neural networks of long-term memory architecture [9] were used. They are effective for the analysis of sequences of various objects, in particular for the recognition of printed and handwritten texts.

Let us consider the comparison of words represented as text lines. Using the distance of Levenshtein [1] $lev_{A,B}(A,B)$ we determine the similarity coefficient for two words as follows

$$coeff_{OCR}(A,B) = 1 - lev_{A,B}(A,B) / max(|A|,|B|),$$

where $max(|A|,|B|)$ is the maximum of the word lengths A and B.

Two words will be considered coordinated, if $coeff_{OCR}(A,B)$ is more than a given threshold $word\_ocr\_simil$.

To compare the recognized words, you must apply the following operations related to OCR Tesseract recognition features:

- ignore case, as one of the common recognition errors is change of the case of random letters within a word;
- identification of some characters, such as short and long dashes, hyphens and minus signs, different types of quotes, as at recognition they can be replaced by each other;
- ignoring punctuation because possible errors are irrelevant when comparing two documents;
- identification of characters that are similar in terms of recognition mechanism, such as the letter O and the digit 0, the letter Z and the digit 3 [11].

The method described above produces a large number of false positives, which can be reduced by another word comparison method based on the comparison of singular points.

Feature points were detected for the model and test images, and RFD descriptors were calculated for the found feature points [5]. Further, an algorithm based on the RANSAC method [4] is used to find the optimal transformation.

The descriptors of the points located in the model image, on the transformed test image were calculated, and then two sets of binary descriptors of points with geometrically close location were compared. After comparison, abatement of outliers was carried out and the similarity coefficient was calculated:

$$coeff_{fp} = 1 - |outliers| / (|outliers| + |inliers|),$$

where $|outliers|$ is the amount of outliers, $|inliers|$ is the amount of inliers. Two words were considered to be coordinated if $coeff_{fp}$ for them is more than a predetermined threshold $word\_fp\_simil$.

The results of combining methods based on recognition and feature points consist in the found word matches of the model and test images. The number of false positives in the comparison results can be reduced by the analysis of words that are neighboring to uncoordinated words.

We checked the sticking of two words and the division of the word into several parts, which are the result of incorrect segmentation. An example is when the bigger part of the split word will be aligned with the model word, and the smaller part will be considered as the inserted word.

The figures below show several cases of false positives. In Fig. 1–6, the words that do not match are marked in red and blue, if the test word is coordinated with the word from the model page, the top word corresponds to the model document, and the bottom word corresponds to the test document.

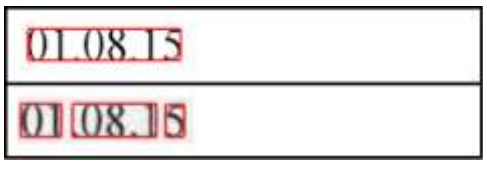


Fig. 1. The test word is divided into parts, the correspondence is not established either for the model word or for parts of the test word

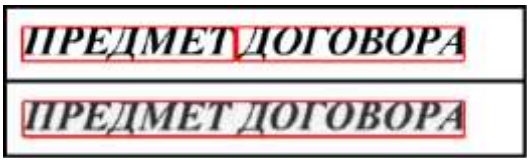


Fig. 2. Two test words are sticked together, the correspondence is not established either for the model word or for parts of the test word

Fig. 3. Two test words are sticked together, the correspondence is not established for the second part of the model word, the other words are coordinated.

Fig. 4. The test word is divided into parts, while the model word and the first part of the test word are coordinated, and the rest of the test word is not matched

Fig. 5. Two test words are sticked together, the correspondence is not established for the first part of the model word, the other words are coordinated

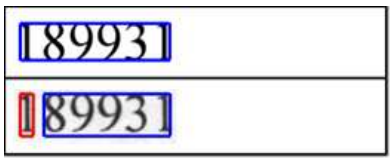


Fig. 6. The test word is divided into parts, while the model word and the second part of the test word are coordinated, and the correspondence to the first part of the test word is not found

In such cases, the coordination of words was clarified by combining or splitting the relevant parts of words.

Also, the check for word shift on the lines was performed (an example is shown in Fig. 6). At modifications like word insertion and deletion, words can be shifted or moved to other lines, which prevents row-by-row comparison, because the shifted word will be considered a modification, as shown in the figure. To avoid such false positives, words line shift check is performed. To do this, when a shift is detected, the next and the previous line are considered and the words on them are compared, regardless of the position on the line.

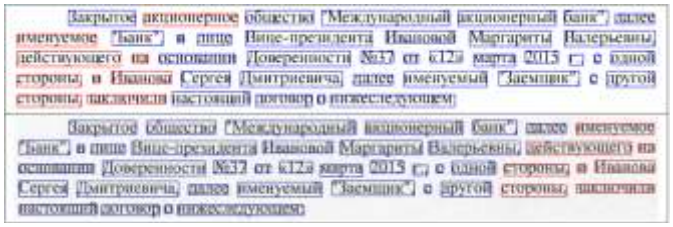


Fig. 7. Example of word shift. The upper part of the figure corresponds to the model document, and the lower part corresponds to the test document. It can be seen that there is a modification of the word “акционерное”, which is why the subsequent words are shifted by lines. The words to which a match is found are denoted with blue, and red denotes the words to which no match is found.

III. Results

For the experiments, we prepared our own test data set consisting of 210 pairs of documents consisting of about seventy thousand words, with 820 modifications, including the insertion and deletion of lines. Half of the test images were scanned at 200 dpi and the other half at 300 dpi. Most of the images did not contain tables and graphic elements.

Testing the algorithm consisted of counting:

- numbers of correctly found modifications (true positive) tp ,
- number of false positives fp ,
- number of real non-found modifications (false negative) fn ,

and determination of the characteristics of accuracy and completeness:

$$Precision = tp / (tp + fp)$$
$$Recall = tp / (tp + fn)$$

Table I presents the results obtained using only OCR Tesseract, Table II – the results of the experiment in combining OCR Tesseract recognition results with OCR Smart Engines.

TABLE I. Results of experiments using OCR Tesseract

	Precision	**Recall**
300 dpi	64,62%	96,36%
200 dpi	64,14%	98,00%

TABLE II. Results of experiments using OCR Tesseract и OCR Smart Engines

	Precision	**Recall**
300 dpi	90,91%	98,18%
200 dpi	74,38%	98,00%

As can be seen from Tables I and II, the use of OCR Smart Engines significantly increases the precision of the modifications: by 10% for images of 200 dpi and 26% for images of 300 dpi, while the recall of the detection of modifications is practically not affected. The final recall of detection of modifications is quite large and is 98% for images of 300 dpi and 200 dpi. Precision is reduced because of false positives arising from recognition errors. The deterioration of Recall accuracy for the 200 dpi case is also due to the increase in the number of recognition errors, examples of which are shown in Fig. 8.

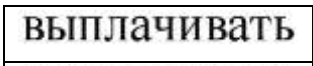
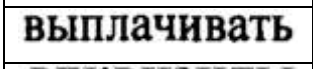
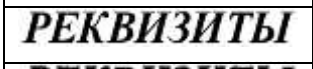
	ВЫПЛАЧИВАТЬ
	ВЫПЛАЧНВАТЬ
	РЕКВИЗИТЬ
	РЕКВИЗИТЬГ

Fig. 8. Examples of typical errors. Images of words are on the left, recognition results are on the right

IV. SUMMARY

The proposed combination of several methods allows to solve the problem of finding discrepancies between scanned copies of business documents. The experiments show that the images scanned with a resolution of 300 dpi achieved a precision of 90.9%, the recall is 98.2%, for images scanned with a resolution of 200 dpi, the precision is 74.4%, and the recall is 98.0%.

V. REFERENCES

[1] Levenshtein V.I. Binary codes with correction of drops, inserts and substitution of symbols. Moscow: *Doklady AN SSSR* [Reports of the USSR Academy of Sciences], vol. 163, № 4, 1965, pp. 845-848. (in Russian)

[2] Slavin O.A. Method of classification of recognized pages of business documents based on the template matching method. *Trudy Sed'moj Mezhdunarodnoj konferencii "Sistemnyj analiz i informacionnye tehnologii" (SAIT – 2017)* [Proc. of the Seventh International conference "System analysis and information technologies" (SAIT – 2017)], 2017, pp. 667-671. (in Russian)

[3] Nikolaev D., Karpenko S., Nikolaev I., Nikolayev P. Hough transform: underestimated tool in the computer vision field. Proc. of the 22th European Conference on Modelling and Simulation, 2008, pp. 238–246.

[4] Fischler M.A. and Bolles R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, comm. Of the ACM 24: 381–395, June 1981, DOI:10.1145/358669.358692.

[5] Shemyakina Y.A., Zhukovsky A.E. and Faradzhev I.A. Investigation of algorithms for computing the projective transformation in the problem of guidance to a planar object from feature points. Moscow: *Iskusstvennyj intellekt i prinjatie reshenij* [Artificial intelligence and decision – making], No. 1, 2017, pp. 43-49. (in Russian)

[6] http://smartengines.biz/smart-id-reader/

[7] Library for recognize the ID cards of individual "Smart IDReader": certificate on state registration of program for computer No. 2016616961 Arlazarov V.V., Nikolaev D.P., Ysilin S.A., Bulatov K.B., Chernov T.S., Slugin D.G., Ilin D. A., Bezmaternykh P.V., Mukovozov A.A., Limonova E.E., No. 2016612014; Appl. 10.03.2016; registered in the register of computer programs 22.06.2016. [1] p.

[8] Chernov T. S., Ilyin D. A., Bezmaternykh P.V., Farajev I.A., Karpenko S.M. Research of Segmentation Methods for Images of Document Textual Blocks Based on the Structural Analysis and Machine Learning. *Vestnik RFFI* [Journal of RFBR], No. 4 (92) October - December 2016, pp. 55-71, DOI: 10.22204/2410-4639-2016-092-04-55-71. (in Russian)

[9] Tang Z., Wang D., Zhang Z. "Recurrent neural network training with dark knowledge transfer," IEEE international conference on acoustics, speech and signal processing (ICASSP) in 2016, pp. 5900-5904.

[10] Slugin D.G., Arlazarov V.V. Search document text fields by using image processing methods. Moscow: *Trudy ISA RAN* [Proc. of ISA RAS], Vol. 67, No. 4. 2017. P. 65-73. (in Russian)

[11] Bulatov B.K., Ilin D.A., Polevoy D.V., Chernyshova Y.S. Problems of recognition machine-readable zones, using small-format digital cameras of mobile devices. *Trudy ISA RAN* [Proc. of ISA RAS], 2015, vol. 65, No. 3, pp. 85-94. (in Russian)