Classification of Recognized Pages of Official Papers Based on the Template Matching Method

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

***Abstract*— The paper deals with the problem of classification of recognized pages of official complex documents, consisting in checking the belonging of a page image to a certain class. To official papers refer documents such as the power of attorney, contract, signature cards and stamps, articles of association, contract, invoice, the certificate of registration, etc. A simple method for classifying official papers is described, which gives acceptable results on the accuracy of the classification necessary for the creation and maintenance of electronic archives of scanned paper documents. The proposed method of classification can be applied in modern CAD for solve problems of analysing the contents of text documents.**

Keywords*— *classification of texts; recognition of documents; OCR; recognition error; template matching

I. Problem description

The task of classification is to establish the correspondence of an undefined object to one or several classes from a predetermined set of classes. The classification algorithm can be built on the basis of machine learning, using some training sample of objects, about which it is known which classes they belong to.

A well-known method is Template Matching, which consists in the preliminary preparation of templates for all possible classes. The decision on whether a test object belongs to a certain class is carried out by the criterion of the minimum (maximum) of some function of the object and its template matching (in work [1] objects-images of symbols are considered). The classification of Template Matching is based on one of the simplest models, at the implementation the basic is the question of representing an object with some template and selecting the mapping function.

To analyse the information contained in the document image (in particular, the classification), you can use the text representation of the document received from the scanned image of the paper document by the OCR program. Currently, one of the most popular solutions is OCR Tesseract, for example, in the work [2] this OCR was chosen to recognize the package of archival documents for the following reasons:

- the possibility of free distribution,
- presentation of recognition results in HOCR (HTML OCR) format with preservation of information about the coordinates of recognized words.

Typical errors of OCR Tesseract can be divided into several types:

- E_1: complete refusal of page recognition,
- E_2: the number of errors is so great that a person cannot understand the document,
- E_3: the page structure is incorrectly recognized,
- E_4: presence of a small number of errors (in incorrectly recognized words there are 1–2 errors).

Let’s consider the task of classifying images of pages of multi-page documents. In the existing flow of images of pages of single-page or multi-page documents, it is necessary to determine the boundaries of each of the available documents and refer these documents to one of the known classes. The resulting page sequences are stored in the electronic archive as separate documents.

The paper describes a simple method for classifying document pages, based on a comparison with pre-prepared class templates.

It is suggested to create class templates based on keywords (phrases) and combining ordered sequences of keywords (phrases).

Currently, several approaches to describing document models are known, for example:

- in the form of a vector model or a "bag of words", i.e. multisets of words entering into the document [3],
- in the form of a language model that takes into account the word order (taking into account the probability of occurrence of a sequence of words) [2].

II. Description of templates and template matching

Let's describe the approach to creating templates using the listed mechanisms and taking into account recognition errors. As the signs will act recognized words in the form of sequences of symbols and its properties: W=($T(W)$, $m_1(W)$, …, $m_6(W)$), where

This work is supported by Russian Foundation for Basic Research (project No. 16-07-01051).

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

- $T(W)$ – *the core of a word, that is, a sequence of symbols and signs* "?" and "*" (symbols"?" and "*" used to specify a set of words with arbitrary characters);
- $m_1(W)$ – distance threshold when *comparing* the template core $T\ (W)$ and the analysed W_r. To compare these two words, it is suggested to use the simplified Levenshtein distance, taking into account only the number of operations of substituting one symbol for another in words of the same length. If $d(T(W), W_r)<m_1(W)$, then the words are considered *identical*, otherwise are different;
- $m_2(W)$ is the restriction on the length of the word W_r when comparing $T(W)$ and $W_{r;}$
- $m_3(W)$ – the dependence on the register;
- $m_4(W)$ – a *word frame* consisting of the coordinates of a rectangle bounding the area on the image of the paper in which the word can be placed;
- $m_5(W)$ – a sign of the *negation of the word*, indicating that the given word should not be in the text.

In the search for a word W in the text of the recognized paper T_r, we check the truth of the following condition:

$$\exists W_r\in T_r{:}d(T(W),W_r)<m_1(W)\wedge(F(W_r)\cap m_4(W)=F(W_r)),$$

where $F(W_r)$ – is a word frame W_r, that is, the word coordinates are extracted from the recognition results in the HOCR format, the abscissas and ordinates are normalized to the width and height of the original image. Generally speaking, several words $\{W_r\}$, identical to the keyword $T(W)$, can be found.

We define the predicate $\mathrm{P}(W,T_r)$=11 for the case $m_5(W)$=0, if at least one word identical to the word W is found in the text of the recognized paper T, and $m_5(W)$=0 otherwise. For words with $m_5(W)$=1, the absence of this word in the text of the recognized document is checked.

Let's define the allocation of words as an ordered set of words $R = W_1,W_2,\ldots,$ for which the presence of each of the words in the recognized paper $T_r$is checked:

$$\mathrm{P}(W_1,\ T_r) \wedge\ \mathrm{P}(W_2,\ T_r) \wedge \ldots \tag{1}$$

and additionally, for each pair of words $W_i\ W_{i+1}$ condition

$$r(W_{i+1}) – r(W_i)< m_6(W), \tag{2}$$

is checked.

Where the function $r(W)$ gives the word number W in the text Tr, assuming that all recognized words are ordered by OCR mechanisms. That is, the parameter $m_6(W)$ determines the distance between adjacent words in the allocation, with$m_6(W_{i+1})=\infty$ only the order of words is checked.

The fulfilment of conditions (1), (2) determines the predicate of the membership of the allocation of R to the text Tr: $\mathrm{P}(R,\ T_r)$=1. In the general case, it requires a search of sets identical to $W_1,W_2\ldots$.

The evaluation of the correspondence to the recognized text Tr of the allocation $d(R,\ T_r)$ is defined as

$$\min(d(W_1,\ T_r),\ d\ (W_2,\ T_r),\ \ldots).$$

We define the *combination* as the set of allocations $S = R_1,R_2,\ldots,$ for which the presence of each of the allocations in the recognized paper T_r is checked:

$$\mathrm{P}(R_1,\ T_r) \wedge\ \mathrm{P}(R_2,\ T_r) \wedge \ldots \tag{3}$$

The order of allocation is not important.

The evaluation of the correspondence to the recognized text T_r of the combination $d(S,\ T_r)$ is defined as

$$\min(d(R_1,\ T_r),\ d\ (R_2,\ T_r),\ \ldots).$$

Finally, we define the *template* M as the set of combinations $S_1,\ S_2,\ \ldots,$ for which the template membership to the recognized text T_r is established by checking the expression

$$\mathrm{P}(M,\ T_r) = \mathrm{P}(S_1,\ T_r) \vee \mathrm{P}(S_2,\ T_r) \vee \ldots \tag{4}$$

The evaluation of the correspondence to the recognized text T_r of the template $d(S,\ T_r)$ is defined as

$$\max(d(S_1,\ T_r),\ d\ (S_2,\ T_r),\ \ldots).$$

In addition to conditions (1), (2), (3), (4), it can be added the check of the hit of words of allocation, combination or template in a certain frame.

To the existing comparisons of the template M with the recognized text T_r, we add tests for the correspondence of some properties of the text (the number of characters on the page, the number of columns of text) with similar properties of the template.

If the set of templates$M_1,\ldots,\ M_n$, is given for n classes, then the task of verifying the correspondence to the class M_i of the recognized page of the document T_r is resulted in calculating the distance $d(M_i,T_r)$ and comparing this distance with the previously known threshold d_1.

The described elements of the templates allow describing keywords and phrases on the assumption that in the text T_r, E_3 and E_4 types recognition errors are possible. To do this, the keywords and phrases are set by masks containing arbitrary single characters and character sequences, and also one-type allocations containing words with characteristic recognition errors are added to the template.

The problem of choosing the best class is solved by calculating the distances $d(M_1,T_r),\ldots,d(M_n,\ T_r)$, by dropping the classes M_j, for which $d(M_j,T_r)>d_1$,, ordering the resulting set in ascending order and preserving one or more alternatives $d(M_{i1},T_r),d(M_{i2},T_r),\ldots$. Conflict resolution $d(M_{i1},\ T_r) = d(M_{i2},\ T_r)$ in the simplest case is resulted in the classification failure, in some cases the conflict is eliminated using additional features.

III. Training Methodology

Let's consider the process of forming templates on a real example for a document flow of 45 classes. The templates were formed in several stages.

A. Stage 1

In the beginning, many *reference documents* were considered. For each class, several samples of ideal documents were prepared, recognized without errors. The textual representations of these documents have been transformed into word bags and phrase bags, i.e. in a multiset of words and phrases from which the *stop-words* were removed. Simple exhaustive algorithms from these multisets were selected single words and phrases from several words inherent with one of the classes. The main attention was paid to the characteristic words from the headings and the title of the sections of the document. In this way, several allocations were selected that separated classes well from others.

B. Stage 2

Further in the selecting documents of small volume, consisting of samples of real documents (single-page and multi-page), the recognition of which gave a wide range of errors of all types E_1 – E_4, pages were selected that required modification of the templates, and pages that could not be classified by the proposed method (to start with, due to a large number of recognition errors). The classification of a selection of q volume was evaluated by the following values:

- n_1 – the number of the first pages of documents that were classified correctly,
- n_2 – the number of the first pages of documents that were classified incorrectly,
- n_3 – the number of the first pages of documents that were not classified,
- k_1 – the number of the not first pages of documents that were not classified,
- k_2 – the number of the not first pages of documents that were classified incorrectly.

Analysis of classification results was carried out using the following criteria:

- *accuracy* $(n_1+k_1)/q$,
- *the fraction of false classification* $(n_2+k_2)/q$,
- *the fraction of refusals from the classification* n_3/q, where $q=(n_1+n_2+n_3+m_1+m_2)$.

C. Stage 3

The first two stages are based on the *use of rules* (templates) for referring to a particular class. For selecting large enough samples (3000 – 30000 samples of each class), it is possible to carry out *machine training, f*or example, by the known method of constructing the CART (Classification and Regression Trees [4]) binary decision tree.

IV. Experiments

For the documents flow consisting of 45 classes, the following classification results were obtained on two test sets:

- consisting of images of papers of medium and poor quality of digitization, obtained for the training stage (884 pages) – n_1=397, n_2=3, m_2=3, n_3=111, m_1=370;
- consisting of images of papers of medium quality of digitization, obtained independently of the training stage (3014 pages) – n_1=832, n_2=8, m_2=1, n_3=146, m_1=2027.

For another flow of single-page documents consisting of 1000 pages of 4 classes, the following characteristics were obtained: n_1=954, n_2=6, n_3=40 (m_1 and m_2 for single-page documents were not counted).

It follows from the data given that the classification method described gives an accuracy of 0.86-0.95, while the false classification does not exceed 0.01, the remaining errors refer to the refusals from classification. That is, the proposed method does not always work, but rarely offers the wrong class.

The accuracy of the method obtained is quite high in comparison with the accuracy of the classification algorithms for web pages described in [5] (the accuracy is 0.75).

The speed of the implemented classification method is quite large: 3000 recognized pages are processed in about 1 minute on an Intel (R) Core (TM) i7-4790 CPU. 3.60 GHz, 16.0 GB, Windows 7 prof 64-bit.

V. Conclusion

The described classification method is simple for self-realization. To reduce the percentage of pages left unclassified, the following steps are possible:

- use of binarization methods to remove background on contaminated documents and documents with a complex background,
- use of more accurate document recognition systems (OCR).

The proposed method allows to classify both single-page and multi-page documents.

The proposed method of classification can be applied in modern CAD, allowing, along with the main function of extracting textual information, to solve problems of analysing the contents of text documents.

Reference

[1] Rafael C. Gonzalez Richard E. Woods Digital Image Processing. 3ND EDITION published by Pearson Education, Inc, publishing as Prentice Hall, Upper Saddle River, New Jersey, 2008. 977 p.

[2] Slavin O.A. Method of classification of recognized pages of business documents on the basis of the method template matching. Proc. 7th International Conference «System analysis and information technology» (SAIT 2017), June 13-18, Svetlogorsk, Russia. 2017. (in Russian)

[3] Martin D., Jurafsky D. Speech and Language Processing. An introduction to natural language processing, computational linguistics, and speech recognition. New Jersey. Pearson Prentice Hall, 2009. 988 p.

[4] Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. Monterey. CA: Wadsworth & Brooks. Cole Advanced Books & Software, 1984. 368 p.

[5] Maslov M.Yu., Palling AA, Trifonov SI. Automatic classification of web sites. RCDL, Dubna, Russia, 2008, p. 230-235. [Electronic resource]. – Access mode http://rcdl2008.jinr.ru/pdf/230_235_paper27.pdf. Date of the application: 04/15/2018. (in Russian)