First Insight into the Processing of the Historical Documents from the Period of Totalitarian Regimes

Lucie Skorkovská[1], Petr Neduchal[2], Zbyněk Zajíc[1], Pavel Ircing[2], Luděk Müller[2],
Lukáš Bureš[2]

[1]University of West Bohemia, Faculty of Applied Sciences, NTIS - New Technologies for the Information Society
Univerzitní 8, 306 14 Plzeň
[2]University of West Bohemia, Faculty of Applied Sciences, NTIS - New Technologies for the Information Society and Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň

lskorkov@ntis.zcu.cz, zzajic@ntis.zcu.cz
{neduchal, ircing, muller, lbures}@ntis.zcu.cz

Abstract. In this paper, we describe the goals and the initial stages of the project ”System for permanent preservation of documentation and presentation of historical sources from the period of totalitarian regimes”. The main goal of this project is to create an integrated archive of the recordings, documents, and photographs that would be accessible online and would provide multifaceted search capabilities (spoken content, biographical information, relevant time period, etc.). The recordings contain retrospect interviews with witnesses of the totalitarian regimes in Czechoslovakia; the other documents are copies of relevant text material and photographs mainly from home archives. This paper focuses on the processing of the historical documents with the optical character recognition and describes the initial experiments.

Keywords: historical sources processing, optical character recognition, document processing

1 Introduction

The main objective of the project ”System for permanent preservation of documentation and presentation of historical sources from the period of totalitarian regimes” is the research and development of software tools for archiving and providing access to the historical resources gathered within the documentary mission of the Institute for the Study of Totalitarian Regimes (USTR)[1].

The Institute for the Study of Totalitarian Regimes studies and impartially evaluates the two times of non-freedom periods of the history of the Czech Republic: the

[1] https://www.ustrcr.cz/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 89-92.

time of the Nazi occupation (1939-1945) and the time of Communist totalitarian power (1948-1989), examines the anti-democratic and criminal activity of state bodies, especially its security services, as well as other organizations based on its ideology. For that purpose, USTR secures and makes accessible to the public the documents related to the time of non-freedom and the time of Communist totalitarian power and converts acquired documents into the electronic form.

Within the documentation activities of the USTR in the years 2008-2015 at least 1000 hours of the audio recordings of the interviews with the witnesses of the totalitarian regimes in Czechoslovakia and 50000 scanned textual documents were created. Nowadays, these documents and recordings are stored on the internal storage of the USTR and are made accessible for the researchers on DVD or through some digital storage services. For only about 160 recordings the text transcription is available, for the rest the researchers must manually go through the whole video recording.

Despite these imperfections, the historic resources gathered in this collection are being used by history experts and researchers from not only the Czech Republic but also from other European countries and USA.

2 Goal of the Project

The main goal of this project is to create an integrated archive of the recordings, searchable documents, and photographs that would be accessible online and would provide multifaceted search capabilities (including the actual spoken content, name and other biographical information, relevant time period, etc.). The archive created in such way would make the work of the researchers more efficient and also would allow a wider scope of interested persons to access these historic resources.

In order to achieve this goal, the methods of automatic speech recognition, automatic indexing, search in recognized recordings, the optical character recognition (OCR) and related techniques of natural language processing will be employed. The rest of the paper describes the first stages of the processing of the scanned documents, which is a challenging task in these circumstances since the source documents are old and of low quality.

3 Optical Character Recognition of the Scanned Typewritten Documents

One of the goals of this project is a transformation of old scans of typewritten documents into searchable text documents. It consists of several tasks that have to be solved. The first one is a development or a selection of an existing Optical Character Recognition (OCR) engine. In the first phase of the project, we have chosen the Google Tesseract OCR engine [3]. The second one is a development of the preprocessing methods that would clear noise and other artifacts in the documents and improve the results of the OCR engine.

There are two additional goals that should also be addressed. The first one is the decomposition of a scanned folder - i.e. group of scans that belong to the same topic or person - into clusters of related documents and the creation of PDF files based on the clusters. The last goal is searching for important meta-information about documents such as its type, title or mentioned persons.

In the first year of the project, we have performed several experiments. In the first experiment, the influence of preprocessing methods on the OCR engine was examined. The second one was focused on the clustering of related documents.

3.1 Preprocessing

During the preprocessing experiment, several methods and their combination were used [4]. One of the most important preprocessing methods is a deskewing (estimation of a skew angle) algorithm. It searches for the rotation that has to be applied in order to get a document with no skew of contained text. A method based on the Fourier Transform was used. The results of this method are promising but it has approximately the same results as the intern deskew algorithm in the Google Tesseract OCR.

We also experimented with the color spaces of the input documents. In the first test, three color variants were compared - particularly red-green-blue (RGB), LAB and gray. The results were approximately 80 % in all cases. In the second phase, the influence of binarization was tested. Results of the binarized documents were slightly better. In the last phase, we propose a different binarization algorithm. Each component of a RGB image is binarized independently. The final image is composed of the binarized components as follows:

$$B_{rgb}(i,j) = \begin{cases} 0 & if\ B_r(i,j) = B_g(i,j) = B_b(i,j) = 0, \\ 1 & otherwise, \end{cases} \tag{1}$$

where $B_{rgb}(i,j)$ is a value of binarized image at coordinates *i,j*. Values B_r, B_g and B_b respectively contained binary value of the image component at coordinates *i,j*. The noise and image artifacts are reduced by this approach. The score of OCR on small dataset was 84% using this approach.

Several other methods were tested during this experiment, particularly histogram equalization [2], [6] and image smoothing algorithms [1]. None of them were significantly successful. All experiments were performed on the small dataset of 25 annotated scans. The score of the used methods was computed using the Levenshtein Distance metric [5].

3.2 Clustering of Consecutive Documents

In the second set of experiments, the decomposition of a document folder into PDF files containing related scans was examined. It is worth mentioning that related scans are usually consecutive. In the first phase, the color similarity was addressed. We have created a feature vector that was composed of the differences of RGB and HSV components computed of the consecutive documents. The K-Nearest Neighbors algorithm (KNN) was trained on the part of a document folder. The test was then performed on two document folders. The first one is the rest of the folder from which KNN was trained. The results are excellent in this case. The second test data was chosen from different document folder. The results are worse in this case - there are clearly non-consecutive scans in one PDF and vice versa. The results seem to be a good starting point for a couple of next experiments. Experiments focused on decomposition based on text extracted by OCR and structural features will be performed in the next phase of the project. But it is not straightforward to pick the right features because consecutive documents can be different in structure and color and vice versa.

4 Conclusion

This paper described the goals of the project ”System for permanent preservation of documentation and presentation of historical sources from the period of totalitarian regimes” and the first stage of the processing of the scanned documents. Based on the results of these first experiments, the subsequent research on the contained text classification and processing can be initiated.

For the OCR experiments, the results were improved by the proposed binarization algorithm. For the scans clustering, it is not straightforward to choose the good features because the consecutive documents can be different in structure and color and vice versa. On the other hand, the results based on the color similarity seem to be a good starting point for further experiments.

References

1. Paris, S., Durand, F.: A fast approximation of the bilateral filter using a signal processing approach. International Journal of Computer Vision 81(1), 24 52
2. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B.T.H., Zimmerman, J.B.: Adaptive histogram equalization and its variations. Comput. Vision Graph. Image Process. 39(3), 355-368
3. Smith, R.: An overview of the tesseract ocr engine. In: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. vol. 2, IEEE (2007), 62 633
4. Sonka, M., Hlavac, V., Boyle, R.: Image processing, analysis, and machine vision. Cengage Learning
5. Yujian, L., Bo, L.: A normalized levenshtein distance metric. IEEE transactions on pattern analysis and machine intelligence 29(6), 109 1095
6. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Graphics Gems IV, Heckbert,P.S. (eds.), Academic Press Professional, Inc., San Diego, CA, USA, 47 485

Acknowledgments: This research was supported by the LINDAT/CLARIN, project of the Ministry of Education of the Czech Republic No. CZ.02.1.01/0.0/0.0/16_013/0001781, by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506 and by the Ministry of Culture Czech Republic, project No.DG16P02B048.