Towards Processing of the Oral History Interviews and Related Printed Documents

Zbyněk Zajíc, Lucie Skorkovská, Petr Neduchal, Pavel Ircing, Josef V. Psutka, Marek Hrúz, Aleš Pražák, Daniel Soutner, Jan Švec, Lukáš Bureš and Luděk Müller

University of West Bohemia, Faculty of Applied Sciences,
NTIS - New Technologies for the Information Society
and Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{zzajic, lskorkov, neduchal, ircing, psutka_j, mhruz, aprazak, dsoutner, honzas, lbures, muller}@ntis.zcu.cz

Abstract

In this paper, we describe the initial stages of our project, the goal of which is to create an integrated archive of the recordings, scanned 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 retrospective interviews with the witnesses of the totalitarian regimes in Czechoslovakia, where the vocabulary used in such interviews consists of many archaic words and named entities that are now quite rare in everyday speech. The scanned documents consist of text materials and photographs mainly from the home archives of the interviewees or the archive of the State Security. These documents are usually typewritten or even handwritten and have really bad optical quality. In order to build an integrated archive, we will employ mainly methods of automatic speech recognition (ASR), automatic indexing and search in recognized recordings and, to a certain extent, also the optical character recognition (OCR). Other natural language processing techniques like topic detection are also planned to be used in the later stages of the project. This paper focuses on the processing of the speech data using ASR and the scanned typewritten documents with OCR and describes the initial experiments.

Keywords: historical sources processing, automatic speech recognition, 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]. This institute studies and impartially evaluates the two totalitarian periods of the history of the Czech Republic: the 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 those periods of suppressed freedom and converts acquired documents into the electronic form.

Within the documentation activities of the USTR, many documents and recordings are stored on the internal storage of the USTR and are made accessible for the researchers on DVDs or through a rudimentary digital storage services. Only a small fraction of the interviews contains the verbatim text transcription – in most of the cases, the interested scholars must manually sift through the whole recording in order to find any information they desire. 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].

The main goal of this project is to create an integrated archive of the recordings, searchable text documents, and photographs that would be accessible online and would provide multifaceted search capabilities (including the actual spoken content, name and other biographical information, the relevant time period, etc.). The archive created in such a 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 the automatic speech recognition (ASR), automatic indexing, search in the stored index , optical character recognition (OCR) and other related techniques of natural language processing will be employed. The rest of the paper describes the first stages of the processing of the interviews with the witnesses of the totalitarian regimes using the ASR system. The witnesses have a specific vocabulary containing archaic words which make the recognition difficult. Further, the first experiment with the automatic processing of the scanned documents is presented, which is a challenging task in these circumstances since the source documents are old and of low quality.

2. The Data

During the years 2008-2015 at least 1000 hours of the audio recordings of the interviews with the witnesses of the totalitarian regimes in Czechoslovakia were created. The structured interview with the witness is ideally conducted

[1] https://www.ustrcr.cz/, http://old.ustrcr.cz/en

[2] http://old.ustrcr.cz/en/international-cooperation

over several sessions. The length of the interview is dependent on the experience of the interviewer and the narrating skills, health condition and life destinies of the interviewee. The word-by-word transcription is currently available for only about 160 audio recordings, the rest is annotated only by the identity of the interviewee.

Furthermore, at least 50 000 paper documents were scanned. These documents are relevant historical text materials and photographs mainly from the home archives of the interviewees or the archive of the State Security. Texts are usually typewritten or even handwritten with poor readability and their scans have really bad optical quality (see examples[3] in Figure 1).

All these data are currently not very well organized. They are just stored in a simple directory structure, each witness's data separated in one folder with subfolders containing audio/video interviews (typically of length in the order of hours, usually without any transcription whatsoever) and scanned relevant documents (photos, investigation files, etc.). The scanned materials almost completely lack the information about their content – some of them are in fact only long uninterrupted sequences of scanned pages from the archive of the State Security, without even the basic description such as which of the pages constitute the same document and where there is boundary between the documents. Therefore our goal is to process these data automatically. First, we plan to use ASR to transcribe and index the content of the audio data. Such automatic processing of the recordings will greatly increase the value of the archive since it will allow the word, phrase and phonetic search in the archive and, as such, will provide researchers a far better orientation in the recorded interviews without the need for costly manual transcription.

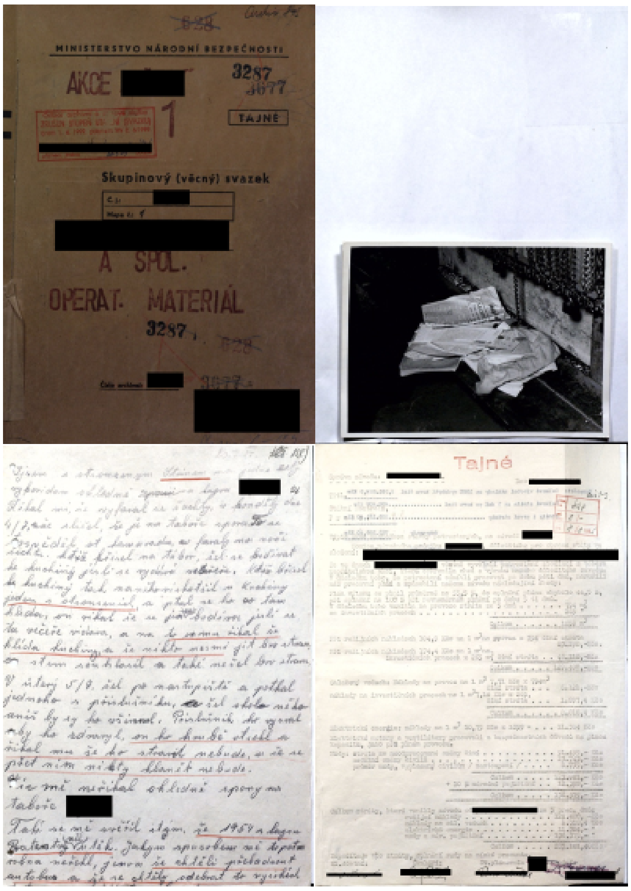

MINISTERSTVO NÁRODNÍ BEZPEČNOSTI
AKCE [redacted]
3287
3677
1
TAJNÉ
Skupinový (věcný) svazek
A SPOL.
OPERAT. MATERIÁL
3287
Tajné

Figure 1: Different examples of the available documents in the database.

Second, the automatic processing of the scanned documents using image processing methods and subsequent natural language processing module will allow for faster and more efficient annotation of individual documents with metadata and will greatly simplify the possible assignment of similarity links among documents, which may be of different media nature (text, audio, or photographs). As a result, the researchers will be able to use a much wider context of the recording.

3. Automatic Speech Recognition of Interviews

From the nature of the speech data, it is difficult to design an ASR system with sufficient accuracy. The speakers are usually elderly, their spontaneous speech is often heavily accented, and because of the nature of the stories, they are often very emotional. The speech quality in individual interviews is however very poor from the ASR point of view (many disfluencies and non-speech events as crying, laughter etc.). The speech was also often affected by using many colloquial words. For the above reasons, it is very difficult to find a suitable data source for training an acoustic model. Due to the great acoustic similarity with the MALACH project, we did the first experiment on the audio recordings with the ASR system designed for the MALACH corpus (Psutka et al., 2014). The MALACH contains 400 randomly selected speakers where only 15 minute segment was transcribed per each speaker. On the Czech part of the corpus, our last results reported in (Švec et al., 2017) achieved a word Accuracy (Acc) of 80.89

3.1. Acoustics models

We used the same acoustics model (AM) as in (Švec et al., 2017) in the experiment. That is, we have followed a typical Kaldi (Povey et al., 2011) training recipe (Kamper et al., 2016) for a Deep-Neural-Network-based AM training. This recipe supports layer-wise RBM pre-training, stochastic gradient descent training supported by GPUs and sequence-discriminative training optimizing sMBR criterion. We have applied the standard 6 layers topology (5 hidden layers, each with 2048 neurons) with a softmax layer. We have used features based on a standard 12-dimensional Cepstral Mean Normalized PLP coefficients with first and second derivatives. The model was trained on 84 hours of MALACH recordings.

3.2. Language Models

Since the vocabulary used in the recordings contains many archaic words and rare named entities, the choice of the training data for the Language Model (LM) is crucial for a good performance of the ASR system.

We have tested two LMs. The first one (denoted "MALACH+") was trained on the MALACH corpus com-

[3] All sensitive data have been blacked out from the example pictures before publishing in this paper.

plemented by relevant text materials from USTR. This additional corpus consists of the few available transcriptions (excluding the ones that correspond to the test set interviews– see below) and other texts on the same topic (books, articles, etc.). In total, the training corpus contains 1.3M tokens and 70k different words.

The second LM was trained only on the available transcriptions of interviews used for testing of the system ("oracle" LM) to show the current performance upper bound of the ASR system. This model was trained using 10 interviews (a total of 60k tokens and 9k different words).

3.3. Results of ASR

As was mentioned previously, the tests were performed on 10 selected interviews, using our ASR system with the two different LM described in Section 3.2. (and the same AM trained on AMALACH corpus). The results can be seen in Table 1.

LM	Acc [%]
MALACH+	58.00
oracle	80.46

Table 1: Results in terms of word accuracy of the ASR system on a small test set (interviews with 10 witnesses) with two different LMs.

The poor result of our initial experiment is most probably caused by the specific language used in the interviews that is strongly connected with Nazi occupation (1939-1945) and the Communist totalitarian state (1948-1989). It seems that more data for the LM creation must be obtained – the easiest way is the manual transcription of the available interviews – in order to significantly improve the ASR performance. The second experiment with the oracle LM indicates the current capabilities of the acoustic model. However, the AM can be adapted using the available recordings to fit the actual acoustics conditions in the interviews.

4. Optical Character Recognition of the Scanned Typewritten Documents

One of the other subtasks of the project is a transformation of the (poor quality) scans of the typewritten documents into the searchable representation. Essentially, our aim is to convert the image data into the electronic text, i.e., perform the Optical Character Recognition (OCR). Note that such a text constitutes the same type of data as the transcripts obtained from ASR in the way described in Section 3. and thus it can be indexed and searched essentially in the same way.

We have decided to use the Google Tesseract OCR engine (Smith, 2007) and concentrate mainly on the appropriate pre-processing (and, in later stages, post-processing) of the image data. In the very first step of the procedure, we needed to automatically classify the available scans into one of the three classes: typewritten (or printed) document, handwritten document and picture. Our plan for the next stage of the project is to train a Convolutional Neural Network (CNN) for classification of each document into these classes. For training this classifier, we need enough reference data. For this purpose, we used an unsupervised clustering method based on CNN (Alexnet) (Krizhevsky et al., 2012) to obtain visual features that are clustered into three clusters using k-means with Euclidean distance. We identified the clusters representing each class by visual inspection. The clusters contained some false data and they need to be manually corrected. Only the documents containing mostly typewritten (printed) text will be processed in the consequent stages.

In order to improve the results of the OCR engine the preprocessing methods that would remove the noise and other artifacts in the documents were examined (see Section 4.1.). Because the documents are stored in a single folder for each witness, without any information about the content in a long uninterrupted sequence of scans, we have proposed a method for the decomposition of a scanned folder – i.e. group of scans that belong to the particular witness – into the clusters of related documents and the creation of PDF files based on the clusters (see Section 4.2.).

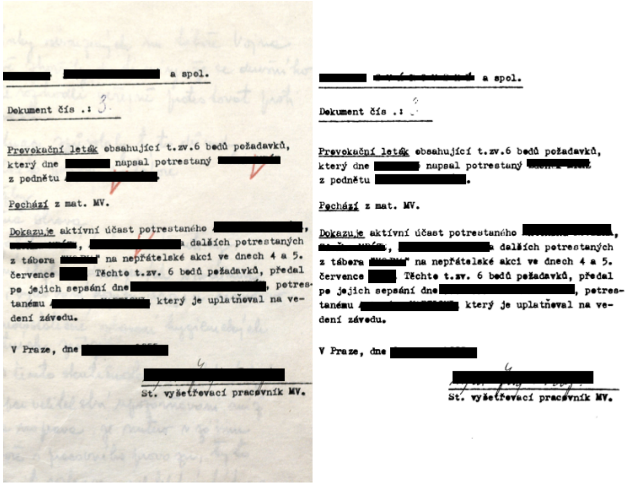
The OCR transcription of these documents allows to search for important meta-information about documents such as its type, title or the mentioned persons which will be used for forming the structure of the final archive. In the first year of the project, we have performed several experiments mainly focused on the influence of the preprocessing methods for the OCR engine and the clustering of the related documents.

4.1. Preprocessing

During the preprocessing experiment, several methods and their combinations were used (Sonka et al., 2008). 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 the contained text. A method based on the Fourier Transform was proposed. This method searches for the skew angle as follows: An input image is transformed into a set of overlapping tiles of a constant size (219 x 248 pixels). Tiles are then sorted in an ascending order with respect to the average value of brightness in the tile. We assume that the tiles with the lowest average brightness value in pixels contain text. The Fast Fourier transform (FFT) is computed over the first eighth tiles in order to get the frequency domain of each tile. An angle is estimated for all frequency domain images. Finally, an average value is computed and the image is rotated by this value. During this process, the interpolation is applied which causes information loss. An example is shown in Figure 2.

The proposed method is promising (see Table 3, line "deskew") but it has worse results than the intern deskew algorithm in the Google Tesseract OCR (line "original RGB image"). Particular deskew algorithm used in Tesseract is unknown for us. Because of the small skew angle in our tested data (smaller than one degree), we deduced that Tesseract is able to process input image without deskewing algorithm – i.e. without information loss. In table 2 we tested our proposed deskew algorithm against the intern one in Tesseract on data with bigger skew angles. In this comparison, our algorithm has much better performance.

Further, we try to extend our preprocessing by binarization method (only with empirically set threshold) for cleaning



[redacted] [redacted] a spol.

Dokument čís .: 3.

Prevokační leták obsahující t.zv.6 bodů požadavků, který dne [redacted] napsal potrestaný [redacted] z podnětu [redacted].

Pochází z mat. MV.

Dokazuje aktívní účast potrestaného [redacted], [redacted], [redacted] a dalších potrestaných z tábora [redacted]" na nepřátelské akci ve dnech 4 a 5. července [redacted] Těchto t.zv. 6 bodů požadavků, předal po jejich sepsání dne [redacted], potrestanému [redacted] který je uplatňoval na vedení závodu.

V Praze, dne [redacted]

[redacted]
St. vyšetřovací pracovník MV.

[redacted] [redacted] a spol.

Dokument čís .:

Prevokační leták obsahující t.zv.6 bodů požadavků, který dne [redacted] napsal potrestaný [redacted] z podnětu [redacted].

Pochází z mat. MV.

Dokazuje aktívní účast potrestaného [redacted], [redacted], [redacted] a dalších potrestaných z tábora [redacted]" na nepřátelské akci ve dnech 4 a 5. července [redacted]. Těchto t.zv. 6 bodů požadavků, předal po jejich sepsání dne [redacted], potrestanému [redacted] který je uplatňoval na vedení závodu.

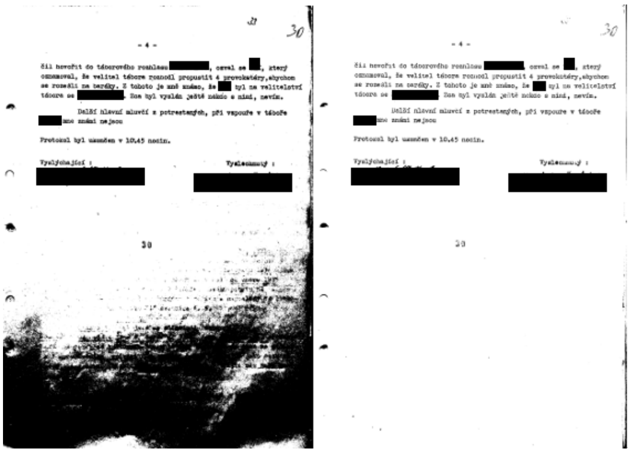
V Praze, dne [redacted]

[redacted]
St. vyšetřovací pracovník MV.

Figure 2: Example of the proposed deskewing approach. Left – the original image, Right – the deskewed and binarized image.

Skew angle approx	Tesseract	Deskew and Tesseract
-4	0.00	84.88
-3	67.13	83.20
-2	64.74	84.15
-1	82.22	86.10
1	85.60	87.25
2	82.77	83.87
3	82.63	84.02
4	78.30	84.76

Table 2: Results in terms of character accuracy of the OCR with our deskew algorithm and with an intern algorithm in Tesseract.



31 30

- 4 -

čil hovořit do táborového rozhlasu [redacted], ozval se [redacted], který oznamoval, že velitel tábora rozhodl propustit 4 provokatéry, abychom se rozešli na baráky. Z tohoto je mně známo, že [redacted] byl na velitelství tábora se [redacted]. Zda byl vyslán ještě někdo s nimi, nevím.

Další hlavní mluvčí z potrestaných, při vzpouře v táboře [redacted] mně známi nejsou

Protokol byl ukončen v 10.45 hodin.

Vyslýchající : [redacted]

Vyslechnutý : [redacted]

30

30

- 4 -

čil hovořit do táborového rozhlasu [redacted], ozval se [redacted], který oznamoval, že velitel tábora rozhodl propustit 4 provokatéry, abychom se rozešli na baráky. Z tohoto je mně známo, že [redacted] byl na velitelství tábora se [redacted]. Zda byl vyslán ještě někdo s nimi, nevím.

Další hlavní mluvčí z potrestaných, při vzpouře v táboře [redacted] mně známi nejsou

Protokol byl ukončen v 10.45 hodin.

Vyslýchající : [redacted]

Vyslechnutý : [redacted]

30

Figure 3: The difference between binarization approaches.

Method	Average accuracy [%]
original RGB image	79.290
deskew	73.378
binary	80.358
binary + deskew	81.838
RGB-binary	**83.753**
RGB-binary + deskew	82.270
bilateral	80.517
bilateral + RGB-binary	80.517
clahe	81.116
clahe + RGB-binary	79.572

Table 3: Results of the preprocessing methods (the average character accuracy of OCR) on the small document's dataset (25 annotated scans containing mostly text).

the noise around the text (line "binary" in Table 3). Next, we propose a different binarization algorithm where each component of an RGB image is binarized independently. The final image is composed of the binarized components as follows:

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

where $B_{rgb}(i,j)$ is a value of the binarized image at coordinates i,j. Values B_r, B_g and B_b respectively contain binary values of the image component at coordinates i,j.

The noise and image artifacts are reduced by this approach more effectively (line "RGB-binary" in Table 3). An example of the difference between the classic grayscale binarization (left) and our proposed method (right) is shown in Figure 3.

Several other methods and their combinations were tested during this experiment, particularly histogram equalization (Pizer et al., 1987) and image smoothing algorithms (Paris and Durand, 2009).

Histogram equalization is based on the method called Contrast Limited Adaptive Histogram Equalization (CLAHE) (Zuiderveld, 1994). This method usually increases the contrast in the image.

Bilateral filtering was used as a representative of the smoothing methods. A bilateral filter is a non-linear, edge-preserving and noise-reducing smoothing filter for images. The intensity value of each pixel in the image is replaced by a weighted average of the intensity values from the nearby pixels. This weight can be based on a Gaussian distribution. Crucially, the weights depend not only on the Euclidean distance of pixels but also on the radiometric differences. This preserves sharp edges by systematically looping through each pixel and adjusting the weights of the adjacent pixels accordingly (Paris and Durand, 2009).

The OCR score with all preprocessing methods was computed using the Levenshtein Distance metric (Yujian and Bo, 2007). It is a character-based evaluation metric that calculates the difference between the two string sequences. The resulting distance is a minimum value of edits - insertions, deletions, and substitutions - that has to be applied in order to transform one string into the other. Based on this distance, the score of the text accuracy can be calculated as follows

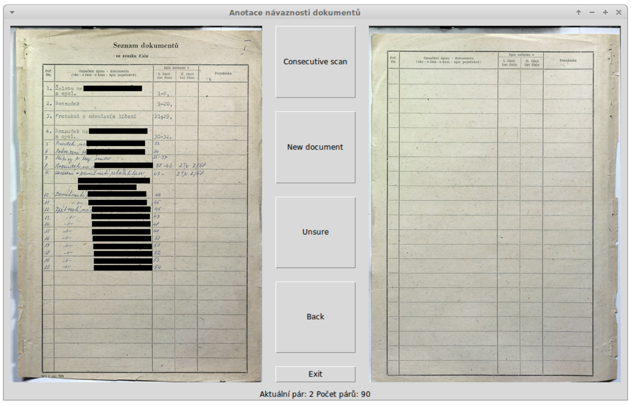
$$Acc = \left(1 - \frac{d}{\max\left(lenght(a), length(b)\right)}\right) \cdot 100[\%], \tag{2}$$

where d is the Levenshtein distance, $length(a), length(b)$ is the length of the input string a and b respectively.

4.2. Cutting of Consecutive Documents

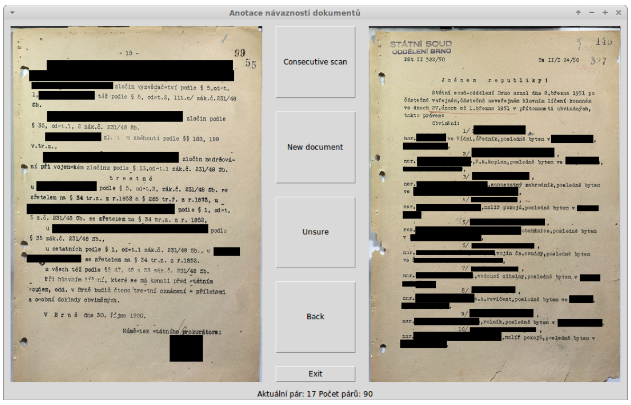
Another task of the project is cutting off the consecutive documents. The goal of this task is to divide individual scans into groups that make up a single document in the document folder. The problem arises from the structure of the available data. All relevant documents for the particular witness are scanned and stored in one folder. Our goal is to preprocess all these scans in one folder and save all the documents as separate pdf – i.e. find the boundaries, where one document ends and the other one begins.

For this purpose, the part of data (1/3 of available data) was manually annotated using developed python application, see Figures 4 and 5. Application windows show two consecutive scans from document folder. The user has to make a decision whether they are connected to the single document or not. Annotated data are then used for training the K-Nearest Neighbors algorithm (KNN) classifier.



Aktuální pár: 2 Počet párů: 90

Figure 4: The window of annotation application with example of consecutive scans.



Aktuální pár: 17 Počet párů: 90

Figure 5: The window of annotation application with example of non-consecutive scans (each from different documents).

The feature vector is composed of the differences between the mean values of color components of a pair of consecutive scans in two color spaces (HSV and RGB). Thus the size of the vector is 6 × 1. This vector is then clustered using the KNN in order to decide about the consecutiveness of documents. Proposed approach exhibits promising results (evaluated only empirically) and seems to be a good starting point for the next experiments. In the future, we want to focus on extending this method by the OCR text-based decomposition and construction of the structural features of the documents. Also, an experiment with a more sophisticated classifier (e.g. Support Vector Machine, Neural Network) will be performed.

5. 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”, the available dataset and the first results of the ASR system and the processing of the scanned documents. Based on these promising results, the subsequent research on the automatic processing of these data to allow the multifaceted search in the dataset is open. Our ASR system used for the transcription of the interviews shows its capabilities for such difficult recordings. For the OCR experiments on the documents, the results were improved by the proposed variant of the preprocessing, necessary for this low quality and high variety of the scanned documents. The proposed methods for the processing of all types of the available data show the first step to fulfill the goal of the project: to create the integrated archive of the recordings and documents that would provide the multifaceted search capabilities.

6. Acknowledgements

This research was supported by the Ministry of Culture Czech Republic, project No.DG16P02B048. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042), is greatly appreciated.

7. Bibliographical References

Kamper, H., Wang, W., and Livescu, K. (2016). Deep convolutional acoustic word embeddings using word-pair side information. In *ICASSP*, pages 4950–4954, oct.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *ImageNet Classification with Deep Convolutional Neural Networks*, pages 1–9.

Paris, S. and Durand, F. (2009). A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach. *International Journal of Computer Vision*, 81(1):24–52.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi Speech Recognition Toolkit. In *Workshop on Automatic Speech Recognition and Understanding*, pages IEEE Catalog No.: CFP11SRW–USB, Hawaii.

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Document Analysis and Recognition*, pages 629–633, Parana.

Sonka, M., Hlavac, V., and Boyle, R. (2008). *Image processing, analysis, and machine vision second edition*. Springer, 3 edition.

Švec, J., Psutka, J. V., Šmídl, L., and Trmal, J. (2017). A relevance score estimation for spoken term detection based on RNN-generated pronunciation embeddings. In *Proceedings of Interspeech 2017*, pages 2934 – 2938, Stockholm.

Yujian, L. and Bo, L. (2007). A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In P.S. Heckbert, editor, *Graphics Gems IV*, chapter VIII. Imag, pages 474–485. Academic Press Professional, Inc., San Diego.

8. Language Resource References

Psutka, Josef and Radová, Vlasta and Ircing, Pavel and Matoušek, Jindřich and Müller, Luděk. (2014). *SC-SFI MALACH Interviews and Transcripts Czech*. Linguistic Data Consortium, USC-SFI MALACH, 1.0, ISLRN 310-213-848-753-5.