Manuscript Investigation in the Sinai II Project

Fabian Hollaus, Ana Camba, Stefan Fiel, Sajid Saleem, Robert Sablatnig

Institute of Computer Aided Automation
Computer Vision Lab
Vienna University of Technology
{holl, acamba, fiel, ssaleem, sab}@caa.tuwien.ac.at

Abstract. This work is concerned with the analysis of historical manuscripts. The manuscripts investigated are partially in a bad condition, due to their age, bad storage conditions etc. These circumstances impede a transcription of the ancient writings, as well as the application of document image analysis methods. Therefore, the writings are imaged with a portable MultiSpectral Imaging acquisition system. By using this non-invasive investigation technique the contrast of the degraded and faded-out writings can be enhanced. In order to gain a further contrast enhancement, a post-processing method has been developed. Additionally, two document analysis methods have been developed in order to facilitate the work of scholars: First, an Optical Character Recognition is described. Second, a method designed for the automated identification of writers of ancient Slavic manuscripts is explained. This paper provides an overview on the manuscript investigation techniques mentioned.

1 Introduction

This work presents image acquisition and processing techniques devoted to the analysis of ancient manuscripts that have been developed in an interdisciplinary project named 'The Enigma of the Sinaitic Glagolitic Tradition' (abbreviated Sinai II). The manuscripts investigated are mainly originating from the 10th to the 12th centuries and are partially in a bad condition, since they contain faded-out writings or are corrupted by background clutter. Additionally, the books imaged contain partially palimpsests, which are writings that have been erased and overwritten. The manuscripts are mainly written in Glagolitic, which is the oldest Slavic script (Miklas, 2004).

Due to the bad condition of the writings, they are partially not legible and thus a transcription by philologist working in our project team is aggravated. Therefore, the writings have been imaged with a portable MultiSpectral Imaging (MSI) system. MSI has proven to be a valuable tool for the acquisition of such ancient manuscripts, since it is capable of capturing information that is invisible to the human eye (Lettner & Sablatnig, 2009).

In this paper our MSI system is introduced and resulting multispectral images are provided in order to demonstrate the capabilities of this non-invasive investigation technique. Such multispectral images are not only used for a transcription by scholars, but form also the basis for further post-processing techniques. In the current work

three of these techniques developed by our project team are presented: First, a method is described in Section 3 that makes use of a dimension reduction technique in order to generate a single image based on a multispectral scan of a document page. Second, a new Optical Character Recognition (OCR) technique for degraded and ancient writings is presented in Section 4. Third, a writer identification method for historical documents is introduced in Section 5. Finally, a conclusion is drawn at the end of this paper.

2 MultiSpectral Imaging

MSI has proven to be capable of enhancing the contrast of degraded or faded-out writings, compared to ordinary white light illuminations – as it has been demonstrated for example in (Easton, Knox, & Christens-Barry, 2003) and (Lettner & Sablatnig, 2009). The works mentioned show that by imaging in selected narrow-band spectral ranges between UltraViolet (UV) and Near InfraRed (IR) the contrast of faded-out writings can be increased, compared to a broadband white light illumination.

Our MSI system allows for an imaging in selected ranges between 365 nm and 1000 nm. 11 different spectral ranges are provided by two multispectral LED panels. Two cameras are used: (1) A Hamamatsu C9300-124 grayscale camera with a spatial resolution of 4000x2672 px and a spectral response between 330 nm and 1000 nm. This camera is used for MSI. (2) A Nikon D4 with a spatial resolution of 4928 x 3280 px. This SLR camera is used for white light images and UV fluorescence photography. One example showing the capabilities of MSI is provided in Figure 1. In this case the ancient writing, which is a palimpsest text, is most visible in the UV fluorescence photograph. It is obvious that the contrast is enhanced in this image, compared to ordinary white light illumination.

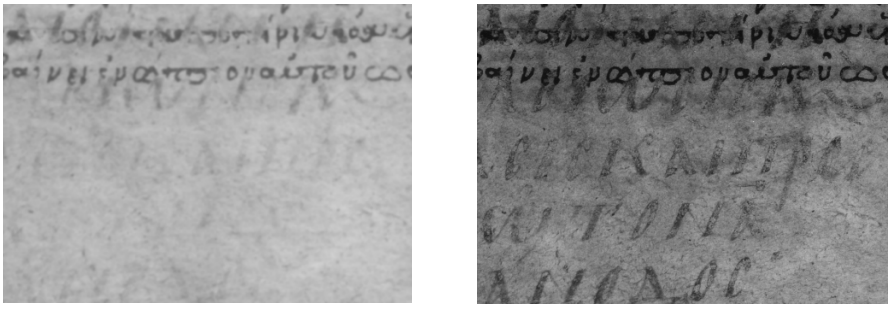

Fig. 1. MSI of palimpsest. (Left) White light image. Two text lines belonging to the younger text are visible in the upper image. (Right) UV fluorescence image, showing the older text.

3 Image Enhancement

While MSI is capable of enhancing the contrast of faded-out writings, it has been shown (Jr., Knox, & Christens-Barry, 2003), that dimension reduction techniques –

like Principal Component Analysis (PCA) – can be used to gain a further contrast enhancement. The dimension reduction techniques are used to lower the third dimension of the multispectral scan in order to extract the relevant information, which is in our case the handwriting. Thus, for manuscripts containing only a single writing, the MSI scan is reduced to just one image emphasizing the ancient text. For palimpsests, on the other hand, the third dimension of the MS scan is reduced to two images emphasizing the two different layers of texts.

In a recent work (Hollaus, Gau, & Sablatnig, 2013), we proposed to make use of a supervised dimension reduction technique named Linear Discriminant Analysis (LDA). Since LDA is a supervised dimension reduction technique it is necessary to label a subset of the multispectral samples as belonging to different classes, whereby in our case the samples are belonging to the foreground or background class. For this purpose we suggested a method that makes use of document analysis methods. More details on the algorithm can be found in (Hollaus et al., 2013). One exemplar output of the algorithm is given in Figure 2. The palimpsest text is again most visible in the UV images, but it is difficult to distinguish from the remaining background. The last image in Figure 2 shows an enhancement result that is gained by combining two LDA resulting images emphasizing the older and the younger text into a pseudo color image. This enhancement result shows that by applying LDA it is not possible to reduce the dimensionality of the multispectral scan, but also to reveal information that is not visible within the multispectral images.

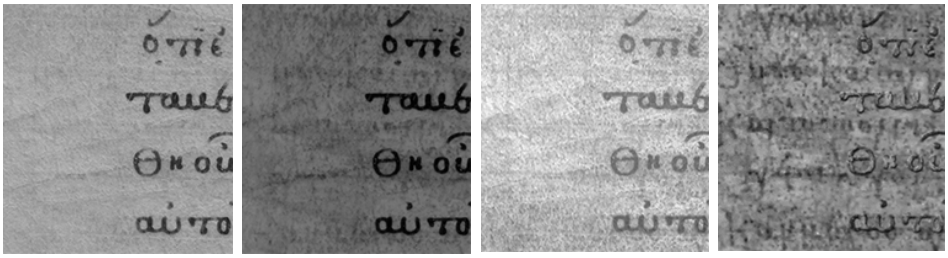

Fig. 2. Enhancement result. (From left to right) White light image. UV fluorescence image. UV reflectography image. Pseudo color image, created by combining enhancement results.

4 Optical Character Recognition

We recently proposed an Optical Character Recognition (OCR) system in (Saleem, Hollaus, & Sablatnig, 2014). The OCR system is especially designed for a strongly degraded Slavonic manuscript, named 'Missale Sinaiticum'. Due to the bad condition of this codex, the document images cannot be binarized successfully in a preprocessing step. Such a document binarization method is used, for example in (Vamvakas, Gatos, Stamatopoulos, & Perantonis, 2008), as a preprocessing step for OCR. Due to the bad condition of the manuscript investigated, we developed a binarization free OCR approach. The method makes use of the Scale Invariant Feature Transform (Lowe, 2004) in order to detect and encode local features. In order to over-

come the problems associated with the detection and the repeatability of the SIFT features, an extension of the SIFT algorithm is used – namely the Dense SIFT approach (van de Sande, Gevers, & Snoek, 2010). In a first step, SIFT features are calculated on a training set consisting of single characters. For each character, a single SIFT feature is calculated. These training features are afterwards used in the recognition stage: The Dense SIFT approach is applied on a given test image. Afterwards, the SIFT features in the training set are compared to the SIFT features found in the test image and the most similar features are used to identify the characters contained in the test image.

The method has been tested on single characters as well as on 15 test panels belonging to different folios of the 'Missale Sinaiticum' manuscript. The method gained an average F-measure of 0.88 on non-degraded test panels and an average F-measure of 0.61 on degraded test panels. In Figure 3 an output of the OCR system is given.

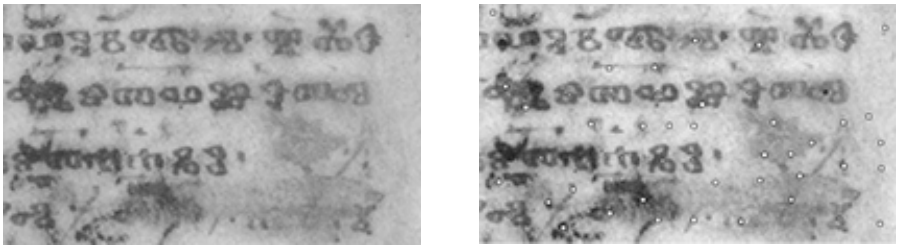

Fig. 3. OCR result. (Left) Input image. (Right) OCR output. Red regions encode falsely classified characters, whereas green depicts correct classifications. Characters that are not colored are not contained in the training set.

5 Writer Identification

Currently, paleographers are performing the task of identifying scribes of writings mainly manually in order to localize, date or authenticate historical writings, (Wolf, Litwak, Dershowitz, Shweka, & Choueka, 2011). A large number of historical documents has been digitized in the past decade and has been made accessible to a growing number of users - for example see (Easton et al., 2003). By automating the task of writer identification, this method can be applied to a vast amount of historical documents and thus become a valuable tool for paleographers.

The historical manuscripts investigated in this work originate from the 10th to 11th centuries and are written in Glagolitic. Writings belonging to five different manuscripts have been examined. The philologists belonging to our project team found that the manuscripts leaves were written by seven scribes.

In order to automatically identify the writers of the ancient scripts, we used a previous work (Fiel & Sablatnig, 2013a) that has been designed for modern Latin writings. Due to the challenges mentioned above, the scribe identification task is complicated compared to modern handwritings. Therefore, we proposed to apply a text re-

gion segmentation method as well as a binarization technique as preprocessing steps (Fiel & Sablatnig, 2013b). These preprocessing steps are used in order to remove background regions and background clutter. It was experimentally found, that such regions containing no text worsen the performance of the algorithm. The writer identification algorithm itself is based on Fisher Kernels (Perronnin, Sánchez, & Mensink, 2010) which are calculated on Visual Vocabularies. The first step is the application of the SIFT algorithm (Lowe, 2004). After the calculation of the SIFT features the visual vocabulary is generated. The interested reader is refereed to (Fiel & Sablatnig, 2013b) for a detailed explanation of the writer identification method proposed.

The algorithm has been tested on 361 images: Each image has been used as an input image and the algorithm returned the most similar image of the 360 remaining images. In 98.9% of the cases considered, the algorithm returned an image that has been written by the same author as the input image. This result and further experiments (Fiel & Sablatnig, 2013b) showed that the method can be successfully applied on the historical manuscripts investigated.

6 Conclusion

This work presents an overview on methods that have been developed for the analysis of historical writings, which are investigated in an interdisciplinary project. Due to the bad condition of the manuscripts, they have been imaged with a portable MSI system. The imaging in selected narrow-band spectral ranges gained a contrast enhancement, which facilitates a transcription by philologists, who belonging to our project team. In order to achieve a further contrast enhancement dimension reduction techniques have been applied: We propose an enhancement method that makes use of spatial and spectral information contained in a multispectral scan. Furthermore, an OCR algorithm has been developed that is especially designed for ancient and degraded methods. The method was successfully applied to a particular Glagolitic manuscript. Additionally, a writer identification method has been applied on 5 different Glagolitic manuscripts. The technique proved to be capable of correctly identifying the majority of the scribes.

References

1. Easton, R. L., Knox, K. T., & Christens-Barry, W. A. (2003). Multispectral Imaging of the Archimedes Palimpsest. In 32nd Applied Image Pattern Recognition Workshop, AIPR 2003 (pp. 111–118). Washington, DC: IEEE Computer Society.
2. Fiel, S., & Sablatnig, R. (2013a). Writer Identification and Writer Retrieval using the Fisher Vector on Visual Vocabularies. In 2013 International Conference on Document Analysis and Recognition (pp. 545–549).
3. Fiel, S., & Sablatnig, R. (2013b). Writer Identification and Writer Retrieval using the Fisher Vector on Visual Vocabularies. In 2013 International Conference on Document Analysis and Recognition (pp. 545–549).
4. Hollaus, F., Gau, M., & Sablatnig, R. (2013). Enhancement of Multispectral Images of Degraded Documents by Employing Spatial Information. In Document Analysis and

Recognition (ICDAR), 2013 12th International Conference on (pp. 145–149). doi:10.1109/ICDAR.2013.36
5. Jr., R. L. E., Knox, K. T., & Christens-Barry, W. A. (2003). Multispectral Imaging of the Archimedes Palimpsest. In AIPR (pp. 111–118).
6. Lettner, M., & Sablatnig, R. (2009). Multispectral Imaging for Analyzing Ancient Manuscripts. In Proc. of 17th European Signal Processing Conference, EUSIPCO 2009 (pp. 1200–1204). Glasgow, Scotland.
7. Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60(2), 91–110.
8. Miklas, H. (2004). Analysis of Traditional Written Sources with the Aid of Modern Technologies. In Conference on Electronic Visualisation and the Arts (EVA).
9. Perronnin, F., Sánchez, J., & Mensink, T. (2010). {I}mproving the fisher kernel for large-scale image classification. In Proceedings of the 11th European Conference on Computer Vision: Part IV (pp. 143–156). Berlin, Heidelberg: Springer-Verlag.
10. Saleem, S., Hollaus, F., & Sablatnig, R. (2014). Recognition of degraded ancient characters based on dense SIFT. In DATeCH (pp. 15–20).
11. Vamvakas, G., Gatos, B., Stamatopoulos, N., & Perantonis, S. J. (2008). A Complete Optical Character Recognition Methodology for Historical Documents. In Document Analysis Systems (pp. 525–532).
12. Van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating Color Descriptors for Object and Scene Recognition. IEEE Trans. Pattern Anal. Mach. Intell., 32(9), 1582–1596.
13. Wolf, L., Litwak, L., Dershowitz, N., Shweka, R., & Choueka, Y. (2011). Active clustering of document fragments using information derived from both images and catalogs. In ICCV (pp. 1661–1667).