Modern Technologies for Manuscript Research

Àâòîð Melanie Gau
10.08.2012 ã.
Ïîñëåäíåå îáíîâëåíèå 16.10.2012 ã.

Summary. This paper presents an overview of image acquisition and post-processing technologies developed in the interdisciplinary project The Enigma of the Sinaitic Glagolitic Tradition. A multi-spectral image recording system using a combination of LED illumination and spectral filtering is described. The possibilities of two different methods of Blind Source Separation, namely Principal Component Analysis (PCA) and Independent Component Analysis (ICA), applied on palimpsest documents are discussed in combination with multispectral input information. We also introduce a new approach to Optical Character Recognition (OCR) that is independent of preceding segmentation of fore- and background, but is based upon local descriptor information. Finally we present a handy and fast image viewer program specialized on multispectral and low contrast images.

Introduction

In the project The Enigma of the Sinaitic Glagolitic Tradition (ASF 19608-G12), the follow-up project of The Sinaitic Glagolitic Sacramentary (Euchologium) Fragments (ASF 23133), an interdisciplinary project group of philologists and computational imaging researchers specialize in developing computational means for digitization, image enhancement, text decipherment, and visualization of historical manuscripts to facilitate philological investigations.

Multispectral Imaging

MSI provides substantial improvement of the legibility of manuscripts that are degraded and barely decipherable under visible light (VIS) (cf. [Knox, 2008]). While the human eye responds to wavelengths ranging between approximately 400 and 700nm, our imaging system allows also for an image acquisition beyond the visible spectrum. Particularly, we make use of three imaging techniques: UltraViolet (UV) reflectography for reflected UV light, whereby the reflected VIS light is filtered out, UV fluorescence, where the incident UV light is filtered out and only the fluorescence in the VIS range is captured, and InfraRed (IR) reflectography, where the VIS part of the light spectrum is filtered out, so that the resulting images exhibit solely the light reflected in the IR range.

Acquisition Setup

We use a Hamamatsu C9300-124 gray scale camera with a spectral response from 330 to 1000 nm. In order to select a certain spectral range, a filter wheel is used that is mounted in front of the camera with 8 different optical filters (cf. Table 1)/Additionally, a Nikon D2Xs SLR camera is used for

RGB and UV fluorescence photographs.

Filter type	Characteristic
SP 400	UV reflectography
LP 400	VIS-IR / UV fluorescence
BP 450	VIS-IR
BP 550	VIS-IR

BP 650	
VIS-IR	
BP 780	
VIS-IR	
LP 800	
IR reflectography	
No filter	
VIS-IR	

Table 1: Optical filters fitted in the filter wheel. SP is a short pass, LP a long pass filter, and BP a band pass filter. The LP 400 filter allows for VIS-IR and UV fluorescence photographs, because white light and UV illumination are used in combination.

The lighting system has improved in the course of the projects. The

first setup included UV and halogen lamps in combination with the optical filters. In the current acquisition setup two LED panels with 13 different narrow spectral bands (cf. Fig. 1) replace the former light sources.

Fig. 1. Spectra of the LED panels

Additionally, four white light
LED panels are used for the RGB photographs, because LED lighting reduces heat
and stress on the manuscripts [Christens-Barry, 2012] and makes subsequent image registration steps[1] obsolete, except for UV reflectography
and fluorescence photographs [Lettner et al., 2007].

The schematic illustration of the current acquisition setup (cf. Fig.
2) shows an object placed on a plate, which is mounted on a linear unit
for automated movement between both cameras. Two diffusers in front of the LED
panels guarantee a uniform light distribution.

Fig. 2. Schematic illustration of the image acquisition
setup

Blind Source Separation

This procedure, the separation of mixed signals, is tailored especially
for palimpsest images and images of degraded manuscripts (cf. the
multi-spectral sample Fig. 3).

Although the underwriting is visible under UV illumination, it is still
barely legible due to low contrast to the background and additional background
noise. Both methods to enhance the older text belong to the category of
blind source separation, which applies statistical assumptions and blind[2] demixing. In our case, the sources to be separated
are the two writing-layers, mold, parchment, etc. The mixtures are given in the
form of multi-spectral images.

The first technique, Principal Component Analysis (PCA), is a
statistical method to transform correlated into uncorrelated variables[3]. The second, Independent Component Analysis (ICA),
other than PCA, transforms statistically independent signals[4] (cf. [Hyvärinen et al., 2001]).

Fig. 3. Palimpsest illuminated at different wavelengths[5]

Experimental results show that contrast enhancement could be
successfully executed by both source separating techniques, but the PCA
approach was often defeated by the ICA approach. Still there are cases, where
the ICA algorithm could not enhance the contrast of the underwriting. In Fig. 4 we can see that the writing in the PCA output (Fig. 4 last two rows, left) has lower background contrast than the ICA output
(last two rows, right). Furthermore, the red initials (first row, left) are visible

in the PCA outputs, whereas the ICA algorithm identified them correctly as a different source.

Figure 4 last two rows, left) has lower background contrast than the ICA output (last two rows, right). Furthermore, the red initials (first row, left) are visible in the PCA outputs, whereas the ICA algorithm identified them correctly as a different source.

Fig. 4. 1st row: Nikon camera; left:: white light; right:: UV fluorescence image. 2nd/3rd row:
left:: PCA, right::
ICA approach

Optical Character Recognition (OCR)

Contrary to other state-of-the-art techniques[6] [Vinciarelli, 2002] the here described OCR system does not require a preceding binarization or character segmentation step, which often is an impossible challenge [Gatos et al., 2006] for low contrast, damaged, or partially faded-out manuscripts (cf. Fig. 5 (b) [Sauvola et al., 2000]). The output of the proposed OCR algorithm
is shown in Fig. 5 (c).

The OCR algorithm is based on local descriptors which describe the local neighborhood around an interest point. Interest points are prominent image points; in the proposed algorithm they are found by applying the Difference-of-Gaussians approach. In order to calculate reliable descriptors, the system makes use of the Scale Invariant Feature Transform (SIFT) [Lowe, 2004]. One advantage of the SIFT approach is its invariance to certain transformations, including translation, rotation, and scaling.

Then the local descriptors are classified by applying a multi-kernel Support Vector Machine (SVM). This SVM is trained by using 20 different training images per character. The output of the classification for each local descriptor is a probability histogram for character classes. Afterwards, characters are localized by clustering the interest points by the k-means clustering algorithm, which assigns each character to a certain cluster. In a final classification step, the class probabilities of each descriptor in a cluster are considered. The probabilities are accumulated in a voting step, assigning the most probable character to the relevant cluster.

Fig. 5. Binarization and OCR applied on a portion of a faded Glagolitic

manuscript: (a) Input image, (b) binarization result of [Sauvola et al., 2000],

(c) characters recognized by the proposed system.

Our experiments were executed on 1055 characters
belonging to 10 different classes which were divided into two different sets:
regular and degraded characters. The results are given in Table 2.

Quantity

Recall

Precision

F0.5-Score

Example

Regular

913

0.732

0.862

0.792

Degraded

142

0.296

0.539

0.382

Table 2: Performance of the proposed OCR system.

It can be seen that the performance gained on degraded, faded-out characters is considerably inferior to the results achieved on regular text (cf. last column of Table 2).

Nomacs Image Lounge Image Viewer

Long-term experiences with multispectral images and low contrast pictures of damaged manuscripts have led to the development of a specialized image viewing program the Nomacs viewer. It is small, fast, able to handle the most common image formats[7] and runs on all major operating systems. This freeware system under GNU Public License v3 [Diem et al., 2011/12] is currently available in three languages: English, German, and Russian.

The Nomacs viewer counters several disadvantages of other programs on the market both for philological investigations, as well as for computational examinations of historical manuscripts. Its key features for the purposes of multispectral image visualization and manuscript including palimpsest research are:

1. Synchronizing multiple instances of an image (e.g. different spectra) including zooming and panning of all synchronized images, as well as synchronized moving to the next/previous file

2. A pseudocolor function for manipulating the color contrast of low contrast writing or palimpsest text (cf. Figure 6).

Fig. 6. High contrast false color image

3. Displaying
of metadata and exif information

4. Fast
thumbnail preview

5. Overlaying
of two or more images (with adjustable opacity)

6. Synchronizing
and sending multiple instances in the LAN

Acknowledgements

The research was funded by the Austrian Science Fund (FWF): P23133. We would like to say special thanks to St. Catherine's Monastery, Mount Sinai, for the kind access to the research objects

References

Christens-Barry, 2012 Christens-Barry, B.: MegaVision Archival and Cultural Heritage Imaging. http://www.mega-vision.com/cultural_heritage.html, 2012.

Diem et al., 2011/12 Diem, M., Fiel, S., Kleber, F.: Nomacs Image Lounge, http://www.nomacs.org, 2011–2012.

Knox, 2008 Knox, K.T. Enhancement of Overwritten Text in the Archimedes Palimpsest // Society of
Photo-Optical Instrumentation Engineers Conference Series, Vol. 6810. 2008.

Lettner et al., 2007 Lettner, M., Diem, M., Sablatnig, R., Miklas H.: Registration of Multispectral Manuscript Images as Prerequisite for Computer Aided Script Description // 12th
Computer Vision Winter Workshop. St. Lambrecht, 2007. Pp. 51–58.

Hyvärinen et al., 2001 Hyvärinen, A., Karhunen, J., Oja, E. Independent Component Analysis, John Wiley & Sons, 2001.

Sauvola et al., 2000 Sauvola, J.J., Pietikäinen, M. Adaptive Document Image Binarization // Pattern
Recognition. 2000. Vol. 33. ¹ 2. Pp.225–236.

Gatos et al., 2006 Gatos B., Pratikakis I., Perantonis, S.J.: Adaptive Degraded Document Image Binarization // Pattern Recognition. 2006. Vol. 39. ¹ 3. Pp.317–327.

Lowe, 2004 Lowe , D.G. Distinctive Image Features from Scale-Invariant Keypoints // International
Journal of Computer Vision. 2004. Vol. 60. ¹ 2. Pp. 91–110.

Vinciarelli, 2002
Vinciarelli A. A Survey on Off-Line Cursive Word Recognition // Pattern Recognition. 2002. Vol. 35. ¹ 7. Pp. 1433–1446.

[1] Since by the use of LED lights the incident light is already filtered, there is no demand for further filtering with optical filters. Only for UV reflectography and fluorescence images distorting wavelengths above – or respectively below – 400nm must still be filtered out. In that case image registration is required.

[2] No additional information about the mixing process is given.

[3] The transformation found is orthogonal and projects
the data in such a manner that the largest variance of the projected data lies
on the first coordinate.

[4] Contrary to the PCA the ICA finds a transformation
that is not necessary orthogonal. One difficulty of the ICA approach results
from the fact that the number of sources has to be defined by the user. We have
noticed that for various manuscript leaves the number of sources has to be
found empirically.

[5] It can be seen that both writings are visible under
UV illumination, but the underwriting vanishes under IR light. Hence it can be
concluded that the ancient text reflects more red light than the younger.

[6] Those techniques are designed for modern
documents and hence they rely on the assumption that the foreground –
background separation can be carried out correctly.

[7] jpg, png, tif, bmp, ppm, xbm, xpm, gif (no
animations), ico (no scales), pbm, pgm, jps, pns; colors are not corrected:
nef, crw, cr2, arw; just main image is displayed: mpo