Digitalisierung von historischen Handschriften mithilfe von Multispektralaufnahmen und Bildverarbeitungstechniken

Digitalization of Ancient Manuscripts with the Aid of Multi-Spectral Imaging and Image Processing Techniques

Fabian Hollaus, Melanie Gau and Robert Sablatnig
Institute of Computer Aided Automation,
Computer Vision Lab
Vienna University of Technology
Favoritenstr. 9/1832, 1040 Vienna
Tel.: +43-1-58801-18382, Fax: +43-1-58801-18399
E-Mail: {holl, mgau, sab}@caa.tuwien.ac.at, Internet: www.caa.tuwien.ac.at/cvl

Zusammenfassung:

Diese Arbeit präsentiert Methoden zur Digitalisierung und Lesbarkeitsverbesserung von historischen Handschriften. Wir befassen uns mit Pergament-Dokumenten aus dem 10. bis 12. Jahrhundert, deren Inhalt zur Bewahrung des kulturellen Erbes abgeschrieben wird. Leider erlitten die Manuskripte im Laufe der Zeit verschiedene Schäden durch schlechte Lagerungsbedingungen, usw., die die Abschrift durch Philologen behindernd. Durch multispektrale Aufnahmetechnik, einer zerstörungsfreien Anwendung zur Verbesserung der Lesbarkeit von kaum sichtbaren Zeichen, wurden die Handschriften mit einem tragbaren Aufnahmesystem digitalisiert. Im Vergleich zu regulärem weißen Licht ermöglicht die Beleuchtung mit bestimmten Wellenlängen eine Kontrasterhöhung ausgeblichener Buchstaben. Dies unterstützt bereits die Arbeit von Philologen, aber es ist noch immer eine manuelle Suche nach relevanten Information in allen multispektralen Aufnahmen nötig und Teile der alten Schriften bleiben unlesbar. Wir verwenden daher mehrere Techniken zur Verschmelzung relevanter Bildinformation in einem multispektralen RGB-Bild. Dieser Fusionsprozess hat zwei Vorteile: Einerseits wird eine Untersuchung des kompletten Datensatzes vermieden und andererseits wird der Kontrast von schwer entzifferbaren Zeichen erhöht. Die resultierenden Bilder zeigen, dass die angewandten Techniken in der Lage sind die Leserlichkeit alter Texte zu vergrößern und so deren Abschrift zu erleichtern.

Abstract:

This work presents digitalization and readability enhancement methods for historical handwritings. We are dealing with parchment documents originating from the 10th to 12th centuries and their contents will be transcribed in order to preserve this cultural heritage. Unfortunately, over time the manuscripts suffered various damages due to bad storage conditions, etc. impeding the transcription by philologists. Since it has been shown that multispectral imaging is a non-invasive exploration technique that allows for an enhancement of hardly legible characters, the manuscripts have been digitalized with a portable multispectral imaging system. Compared to regular white light, the illumination with certain wavelengths enables contrast enhancement of faded-out characters. This already supports the work of philologists, but the scholars still have to search manually for relevant information in the entire multispectral scan and parts of the ancient writings still remain undecipherable. Thus, we applied several techniques in order to fuse the relevant information contained in a multispectral scan into a RGB image. This fusion process has two advantages: On the one hand, an investigation of the entire scan is avoided and on the other hand the contrast of the degraded characters is enhanced. Resulting images reveal that the techniques

investigated are capable of increasing the legibility of the ancient texts and thus facilitate the transcription.

1. Introduction:

Starting from the early 1960s multispectral imaging has been successfully used in remote sensing applications (Govender et al. 2007) and has recently proven its usefulness for the non-invasive examination of ancient and degraded handwritings. For example, Easton et al. (Easton et al. 2003) have applied this imaging technique for the investigation of the famous Archimedes palimpsest. A palimpsest is an ancient manuscript, in which an original writing has been tried to erase, e.g. washed or scraped off, and was then overwritten by a younger text. Since parchment was a precious material, this was a common praxis (Easton et al. 2003) and many important texts have only been preserved today as palimpsest underwritings.

This work deals with the imaging of a palimpsest stored in the National Library of Sofia, which contains a Cyrillic overwriting and partially Greek, partially Cyrillic underwritings. While the overwriting is visible under normal tungsten illumination, the underwriting is hardly legible under broadband white light and most visible under certain - see Section 3 - narrow spectral ranges. The object has been imaged with a portable MSI system, designed for the digitalization of historical writings. For each manuscript leaf, a multispectral scan is taken that consists of 16 different photographs taken under varying illumination conditions.

Since the data in such a multispectral scan is highly correlated, we applied several post-processing techniques in order to fuse relevant information into an RGB image. A simple but yet effective technique (Easton et al. 2003) is a pseudocolor image approach, whereby entire bands of a multispectral scan are copied into the channels of a pseudocolor image. A more sophisticated complex of techniques is called spectral unmixing: Such methods aim at separating diverse sources - e.g. a text written with one particular ink - by applying statistical approaches. We applied two source separation techniques, namely Principal Component Analysis (PCA) and Independent Component Analysis (ICA), and compared their performance to the performance of the manual band selection approach by providing representative resulting images.

This work is structured as follows: In the next section other MSI systems and enhancement techniques are presented. Then, the MSI system utilized is detailed in Section 3. Section 4 contains an introduction to the source separation methods used and results are given in Section 5. A summary and outlook complete the paper.

2. Related Work:

Lettner et al. (Lettner et al. 2008) utilized an MSI system, in which the reflected broadband light is filtered with 8 different optical filters. Another imaging setup is introduced in (Rapantzikos and Balas 2005), where 34 spectral bands are provided by optical filters. Easton et al. (Easton et al. 2003) make use of another strategy for obtaining spectral bands: Instead of filtering the reflected light, the incident light is already filtered by using narrowband LEDs. Thus, the heat put on the manuscripts can be reduced compared to tungsten illumination. This illumination system has been used for the imaging of the Archimedes palimpsest and our MSI system makes also usage of such multispectral LEDs.

The aforementioned approaches are also developing readability enhancement techniques. (Lettner et al. 2008) proposed a method based on an extended version of PCA that is applied in order to enhance the contrast of a single text. Contrary, in (Easton et al. 2003) PCA is applied on a multispectral scan in order to separate the different inks found in the Archimedes palimpsest. The same palimpsest is also investigated in (Tonazzini et al. 2007), where PCA and ICA are used for the separation and contrast enhancement of the diverse writings.

3. Multispectral Imaging Setup:

Our MSI system consists of two cameras (cf. Figure 1): The camera on the left is a Hamamatsu NIR grayscale camera with a spatial resolution of 4000 x 2672 px. It has a spectral response between 300 nm and 1000 nm. The camera on the right is a Nikon D2x DSL camera, with a resolution of 4288 x 2848 px. The Nikon camera is used for RGB photographs taken under white light and additionally for UV fluorescence images, whereas the Hamamatsu photographs are taken under all spectral ranges provided by the lighting system. The manuscripts are put on a plate that is mounted on a linear unit. Thus, it is possible to automatically move the manuscripts under both cameras.

The illumination is provided by two Eureka!Light^{TM} (Equipoise imaging, Archimedes project (Easton et al. 2003)) LED panels, which enable an imaging in 11 different narrow spectral ranges. The spectra of the panels are depicted in Figure 2, whereby the range of the visible spectrum is shown at the top of the figure. Compared to a tungsten illumination, the LED illumination reduces the heat on the manuscripts (Christens-Barry 2012) and a filtering of the reflected light with optical filters is not necessary. Hence, distortions arising from the filtering process are avoided. Nevertheless, we still use two optical filters built in a filter wheel - see Figure 1 - for UV reflectography and UV fluorescence photographs in order to cut off the reflected light below and respectively above 400 nm. These images and the Nikon photographs are registered on the remaining Hamamatsu images with an image registration algorithm proposed in (Lettner et al. 2008).

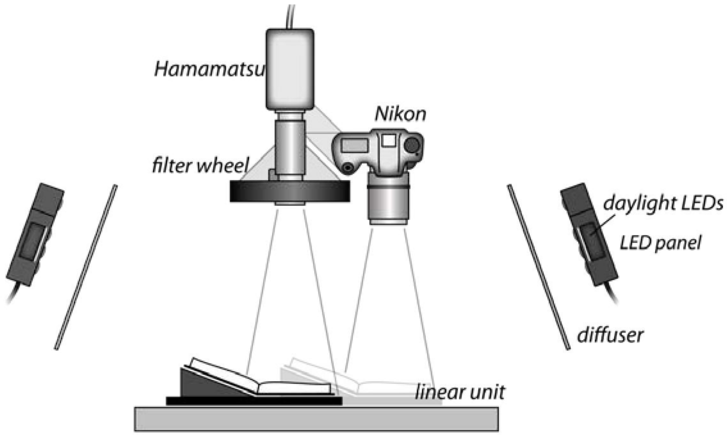


Figure 1: Imaging setup.

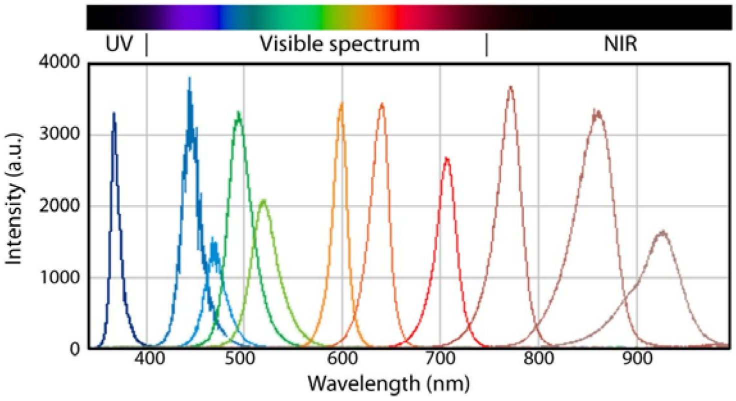
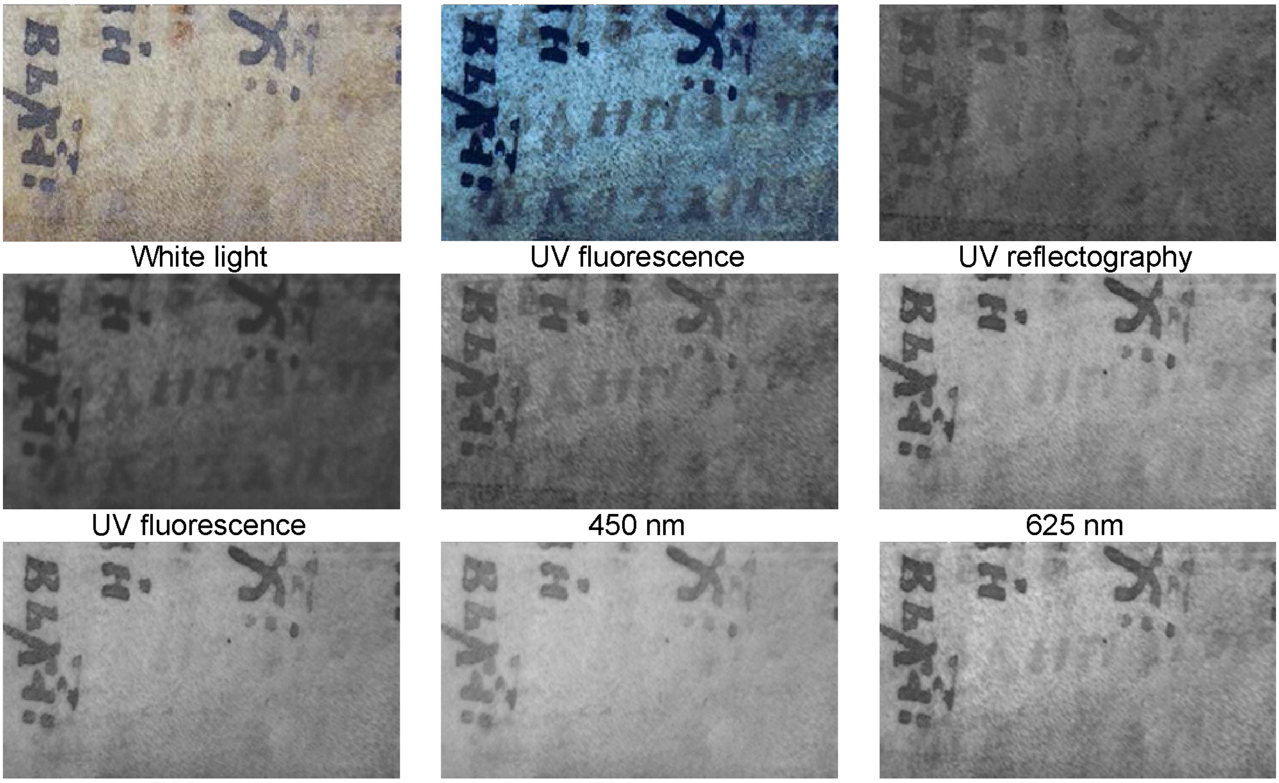


Figure 2: Spectra of the Eureka!Light ^{TM} LED panels

An example for the MSI is shown in Figure 3. It shows 7 images taken with the Hamamatsu camera and two images from the Nikon SLR camera. It can be seen that the horizontally written palimpsest text is only partially visible under white light, whereas it is best recognizable in UV fluorescence images due to the fluorescence of the parchment (Easton et al. 2011). The overwriting is even visible in the NIR range, while the underwriting is completely invisible at this spectral range.



White light UV fluorescence UV reflectography

UV fluorescence 450 nm 625 nm

700 nm 780 nm White light

Figure 3: Portion of the manuscripts imaged under different wavelengths.

4. Image Fusion:

In order to provide images showing both under- and overwriting we applied three different approaches. The first approach utilized is a simple pseudocolor technique as proposed in (Easton et al. 2003): Therein, Easton et al. state that the red channel of a tungsten illuminated photography exhibits mainly the overwriting in the Archimedes palimpsest, whereas the underwriting is most visible in the blue channel of a UV fluorescence image.

Since we made a similar observation we applied the same pseudocolor approach on the palimpsest investigated. Such a pseudocolor image is built up by copying the red channel of a white light image into the red channel of the resulting image. The blue channel of an UV fluorescence image is copied into the green and blue channel of the pseudocolor image. We also used this approach for the fusion of the spectral unmixing results: Therefore, the image showing the overwriting is put into the red channel of a pseudocolor image and the underwriting image is copied into the green and blue channel.

Blind Source Separation:

The applied Blind Source Separation (BSS) techniques are based on a linear [mixing] model, where the multispectral scan $x(t) = [x_1(t), ..., x_n(t)]$ is the result of an unknown mixing process:

$$x(t) = As(t)$$

where A is the unknown mixing matrix and s(t) = [s_1(t), ..., s_m(t)] are the sources; in our case the diverse writings are sources. We assume that m ≤ n, meaning that the number of sources is less or equal than the number of variables. Source separation techniques aim at estimating the unknown source signals by determining the unmixing matrix W that fulfills:

$$y(t) = Wx(t)$$

where $y(t)$ contains tje estimated sources. It is notable that this is a simplified model, since noise is neglected in this definition. Two linear source separation methods have been investigated and are explained in the following.

PCA is a statistically based technique that transforms correlated (normalized) input data into linearly uncorrelated output data. It finds an orthogonal basis in the data and projects the data on the basis found. After subtracting the mean from the data, the eigenvalue decomposition of the covariance matrix $Cov(x)$ is performed:

$$D = V\, Cov(x)V^T,$$

where D is a diagonal matrix and V is orthogonal. Given this variables the demixing matrix is defined by:

$$W = D^{-\frac{1}{2}}V$$

ICA is another strategy for BSS. Its main assumption is that the sources are statistically independent. Two random variables y_1 and y_2 are said to be independent if their joint probability densities $p(y_1)$ and $p(y_2)$ fulfill:

$$p(y_1, y_2) = p(y_1)p(y_{2)})$$

Another assumption of the ICA model is that the random variables have a non-gaussian distribution. ICA models determine the independent components by maximizing the non-Gaussianity of $Wx(t)$. Several strategies for measuring non-Gaussianity and different ICA algorithms have been proposed. We used the FastICA fixed-point algorithm (Hyvärinen and Oja 2000), which uses the negentropy as a measurement for non-Gaussianity. Before the algorithm is applied the data should be preprocessed: Therefore the data is whitened with the PCA approach. In our application, we noticed that the performance of the ICA algorithm is improved, if the dimensionality of the input data is additionally reduced with the PCA approach. The ICA results, which are presented in the following section, have all been produced on a dataset that was at first reduced from 16 to 5 dimensions.

5. Results:

The first two rows in Figure 4 show a portion of the palimpsest imaged under white light and under UV light. The remaining rows in Figure 4 are results of the BSS techniques, whereby the results of the PCA approach are given in the left column and the ICA results are shown in the right column. It can be seen that the ICA results are more accurately, since the diverse inks - i.e. the underwriting, the black overwriting and the red initials belonging to the overwriting - are separated in a more satisfying way. Although the underwriting is most visible in UV fluorescence photographs (compared to the remaining images in the multispectral scan), the contrast of the palimpsest text is still limited. If we compare the UV fluorescence image with the ICA output it is obvious that the ICA technique enhanced the contrast of the ancient writing.

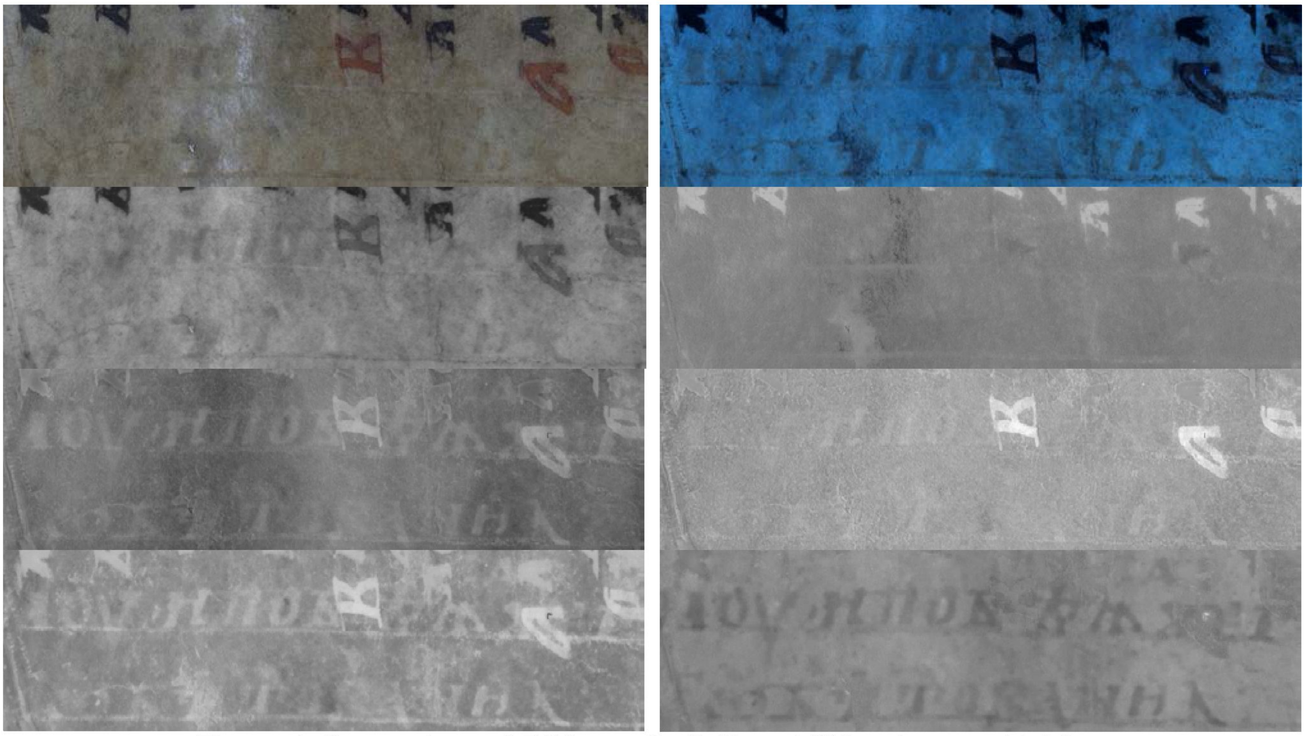


Figure 4: Comparison of PCA results (left) and ICA outcomes (right).

Figure 5 contains an UV fluorescence image and three corresponding pseudocolor images: The right image in the first row was produced with the manual band selection approach proposed in (Easton et al. 2003). The second row contains pseudocolor images that were produced with PCA outputs on the left and ICA results on the right. Compared to the pseudocolor image produced from unprocessed photographs, the underwritten text in the BSS results has a higher contrast. The overwriting is instead most visible in the first row of Figure 5, since the text is contained in all channels, whereas it is only present in the first channel of the BSS pseudocolor images.

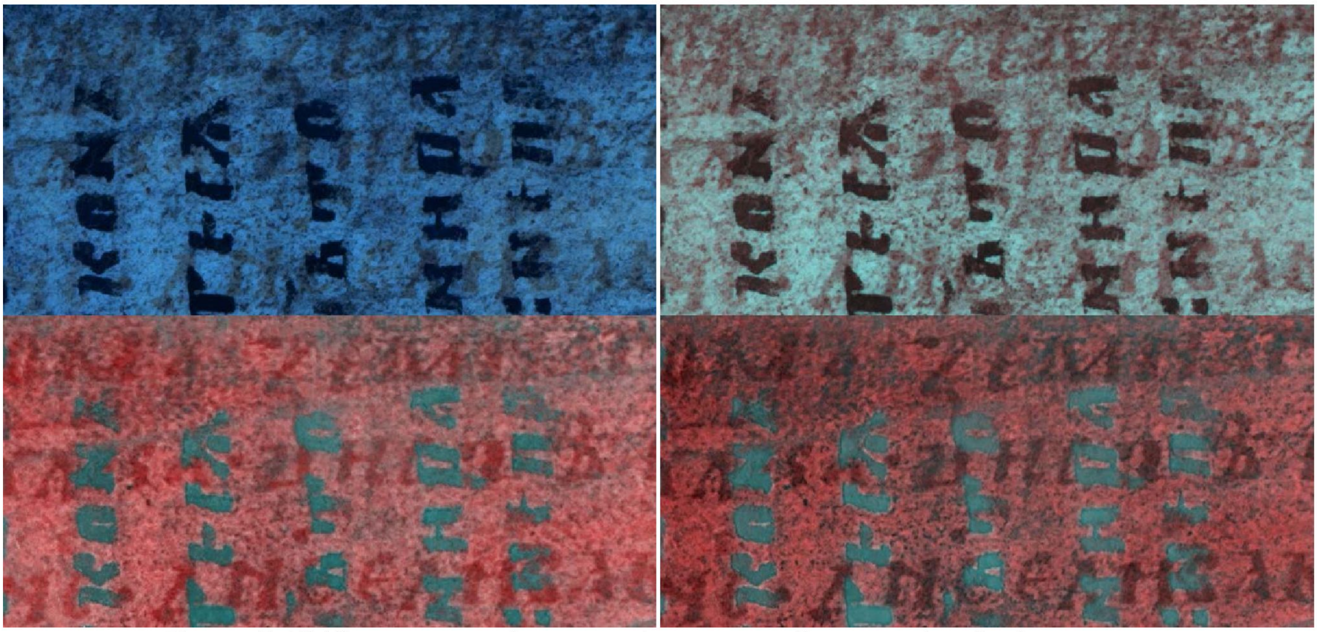


Figure 5: UV fluorescence image and corresponding pseudocolor images.

6. Conclusion:

This paper presents the benefits of MSI and enhancement techniques applied on historic documents. Due to the bad condition of the objects, regular RGB photographs are not sufficient for text transcription . Hence, the writings have been imaged in 11 different spectral ranges and we noticed that the underwritten text is most visible in UV fluorescence images. Nevertheless, also in these images certain characters have a low contrast in comparison to the background. Hence we applied two BSS techniques in order to increase the contrast. It was found that the ICA algorithm is superior compared to PCA technique. In order to make the palimpsest and the overwritten, younger text visible in a single image the separated images containing the under- and overwriting have been fused into pseudocolor images. Additionally, two bands of the multispectral scan have been fused into single RGB images. We believe that such images facilitate the work by philologists, since they show both writings at a glance.

Acknowledgement:

The research was funded by the Austrian Science Fund (FWF): P23133.

References:

Christens-Barry, W.A.: LED Imaging of the Archimedes palimpsest (accessed 2012) http://archimedespalimpsest.org/imaging experimental3.html.

Easton, R., Christens-Barry, W., Knox, K.: Spectral Image Processing and Analysis of the Archimedes Palimpsest. In: 19th European Signal Processing Conference (EUSIPCO 2011). (2011)

Easton, R., Knox, K., Christens-Barry, W.: Multispectral Imaging of the Archimedes Palimpsest. In: 32nd Applied Image Pattern Recognition Workshop, AIPR 2003, Washington, DC, IEEE Computer Society (October 2003) 111--118

Govender M., Chetty K., Bulcock H.: A review of hyperspectral remote sensing and its application in vegetation and water resource studies. Water SA 33 (2) 145--152

Hyärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13(4-5) (2000) 411--430

Lettner, M., Diem, M., Sablatnig, R., Miklas., H.: Registration of Multispectral Manuscript Images as Prerequisite for Computer Aided Script Description. In: 12th Computer Vision Winter Workshop, St.Lambrecht, Austria (2007)

Lettner, M., Diem, M., Sablatnig, R., Miklas., H.: Registration and Enhancing of Multispectral Manuscript Images. In: 16th European Signal Processing Conference (EUSIPCO08). (2008)

Rapantzikos, K., Balas, C.: Hyperspectral imaging: potential in non-destructive analysis of palimpsests. IEEE International Conference on Image Processing 2 (11-14 Sept. 2005) II--618--21

Salerno, E., Tonazzini, A., Bedini, L.: Digital image analysis to enhance underwritten text in the Archimedes palimpsest. International Journal on Document Analysis and Recognition 9(2) (2007) 79--87