Publishing Information

CONTENTS

Article

Image Processing Software for the Recovery of Erased or Damaged Text

Keith T. Knox | Kihei, Hawaii

1. Abstract

An imaging processing software package is being developed to recover erased or damaged text from multispectral images of ancient documents written on parchment or paper. The software is being written in the Java programming language to make it portable to many different computer platforms. The goal of the software project is to make this package of image processing routines available for use anywhere in the world by researchers, students, and potentially even scholars. The architecture of the software has been designed to make it modular, easily expanded, and easy-to-use with an intuitive graphical user interface. The capabilities of the software are demonstrated with examples of recovered text from manuscripts from the library of the Holy Monastery of St Catherine at Mount Sinai in Egypt.

2. Multispectral Imaging

The multispectral imaging system used in this project at St Catherine's Monastery was developed by MegaVision. Called the EurekaVision system, it includes a 50 mega-pixel panchromatic camera, and two panels of LEDs (light-emitting diodes) with diffusers, at several wavelengths across the spectrum from the ultraviolet to the near infrared.[1]
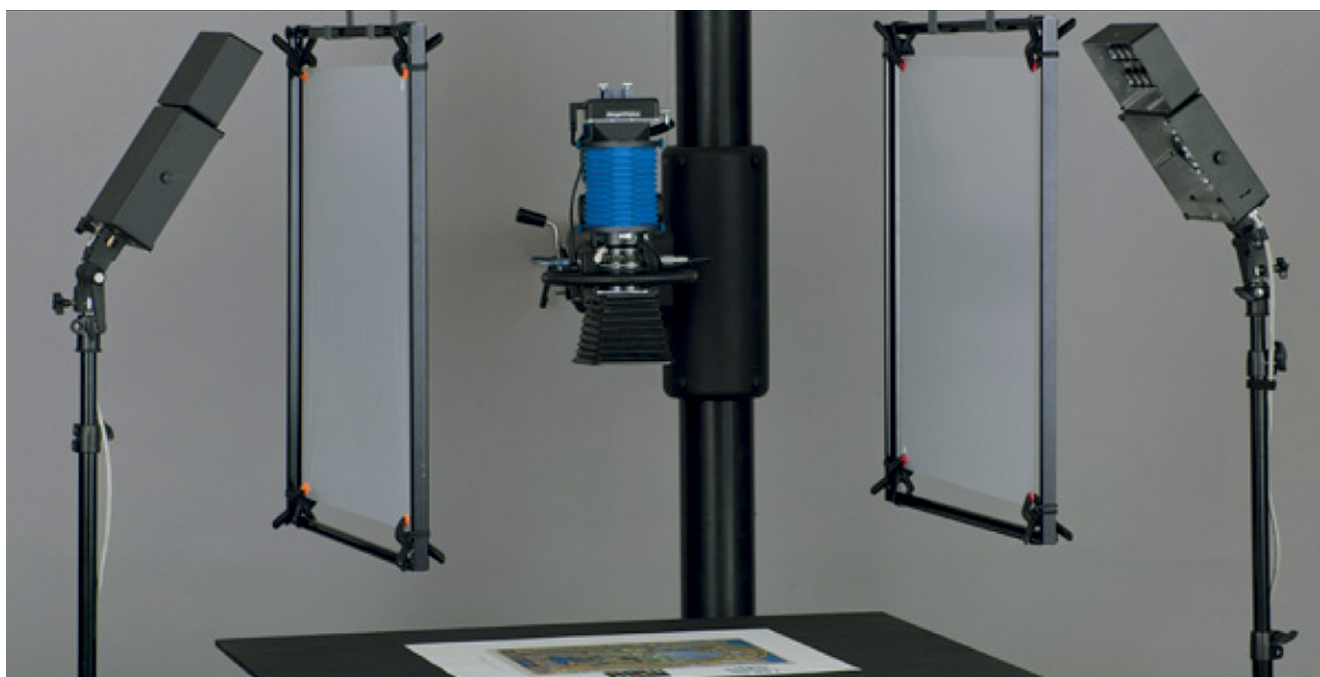

Fig. 1: EurekaVision cultural heritage imaging system from MegaVision.

---

[1] *MegaVision Archival and Cultural Heritage Imaging* <http://www.mega-vision.com/cultural_heritage.html>.
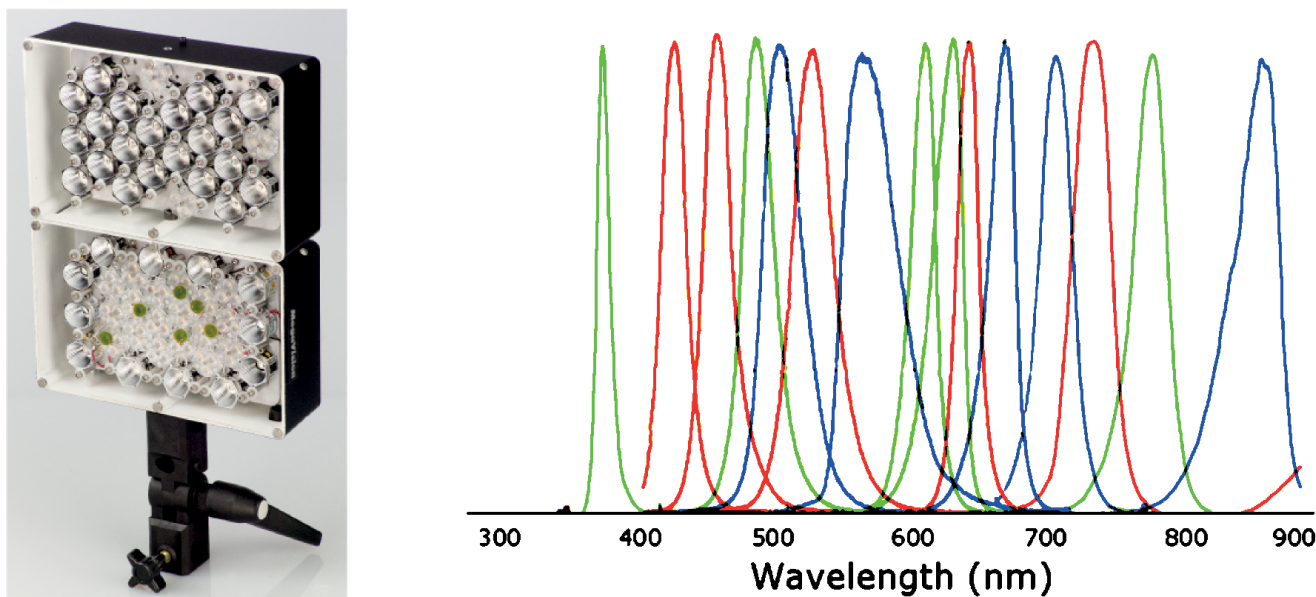
Fig. 2: A light panel from the EurekaVision system with several LEDs and their corresponding spectra.

A MegaVision imaging system is shown in Fig. 1. An early version of this system was used to image the Archimedes Palimpsest.[2] There are two light panels at either side of the copy stand, with diffusers suspended in front of the lights to spread the light uniformly across the manuscript. The LEDs are narrow-band light sources at specific wavelengths across the spectrum from the ultraviolet, through the visible region, and into the near infrared. A single LED panel, with the accompanying spectra of the LEDs, is shown in Fig. 2. During the course of the Sinai Palimpsests Project, LEDs of 12 different wavelengths were used. Since then, the number of wavelengths has been expanded to 15, as shown in Fig. 2.

To image the manuscript and gather the multispectral data, the manuscript is illuminated with light from the LEDs of individual wavelengths. During a single exposure, only the light from LEDs of one wavelength is used. Subsequent exposures capture the response of the manuscript from each of the different wavelengths sequentially. Given that the manuscript and the imaging systems (lights and camera) do not move from exposure to exposure, then all of the multispectral images are registered with respect to each other. This is achievable, because there is no heat generated by the LEDs that might cause the manuscript page to expand or contract between exposures.

In addition to these wavelengths, an option to capture the color variation of the fluorescence is also available.

Typically, due to exposure with ultraviolet light, the substrate of the leaf fluoresces, while the inks do not. This increases the contrast of the writing against the background of the leaf. When the leaf fluoresces, it absorbs the ultraviolet light and re-emits light at lower wavelengths, typically in the visible region. A series of filters are rotated into the light path to filter the fluorescence and record its color variation. Most of these filters are very thin films, and as a result, there is little to no translation of the images due to the thickness of the filters and an arbitrary tilt of the filter within the light path. This maintains registration of the fluorescence-filtered images with the rest of the image set.

Further study of targets specifically designed to test the registration has shown that slight variations do exist in the captured imagery. Although this study has only just been started, initial results show that there are some very slight translations and magnifications induced by the very thin Wratten filters. There are also two glass filters used, one to block and one to pass the reflected ultraviolet light. These thicker filters do induce a noticeable magnification to the images, which needs to be corrected to avoid introducing edge effects in processed results that use those two images. It is beyond the scope of this paper to detail these effects or their mitigations, which will be addressed in future publications.

The spectral response of the parchment and inks can be seen in Fig. 3. The contrast of the images has been adjusted so that the parchment appears to be the same across the

---

[2] Easton, Knox, et al. 2010.
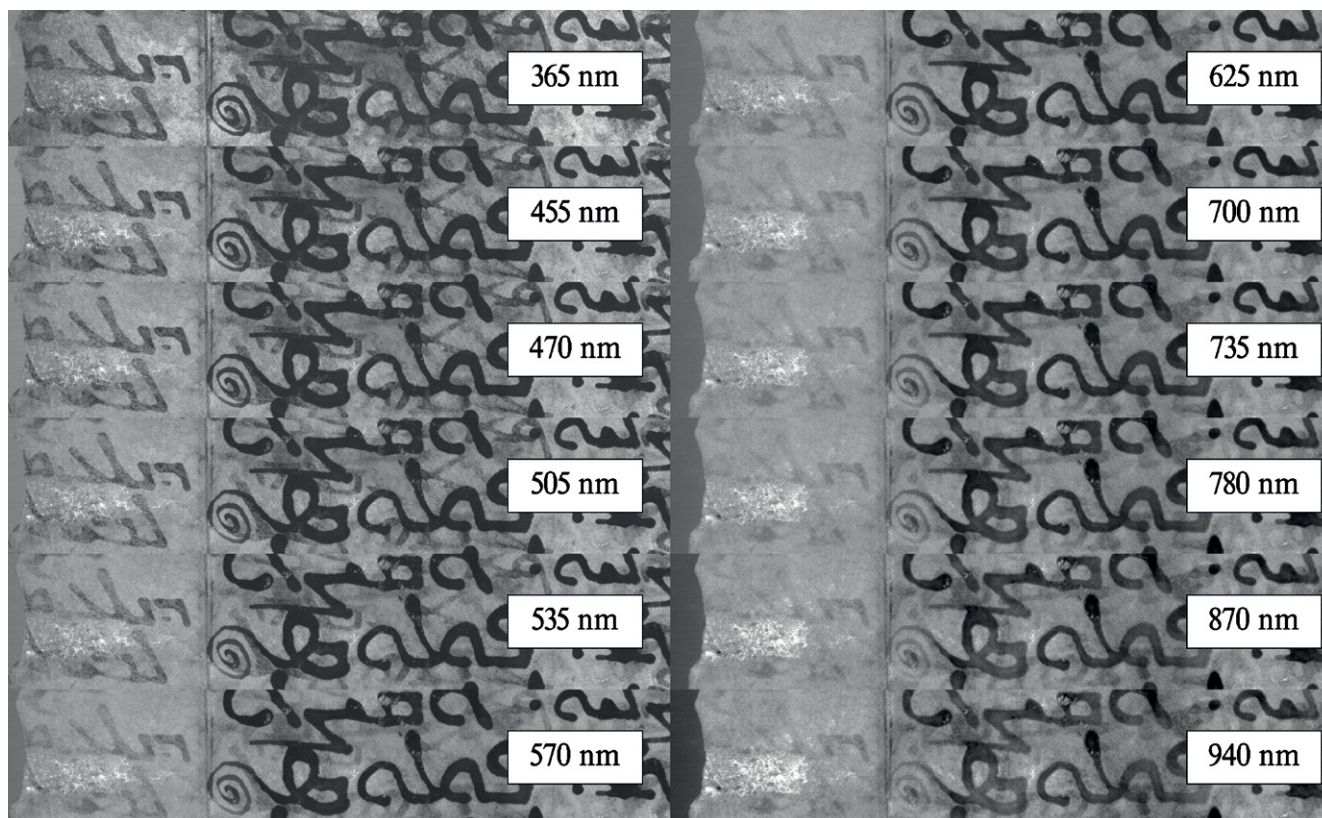
Fig. 3: Several images at different wavelengths of a manuscript parchment containing erased and overwritten handwriting.

wavelengths. Of course, the actual reflectance of the parchment varies across the spectrum. With this contrast adjustment, though, one can see that, in relation to the parchment, the erased ink stains in the parchment (seen on the left in each image) fade rapidly and practically disappear in the near infrared. On the other hand, the overwriting (seen on the right side of each image) maintains high contrast across the spectrum, although it does fade a little bit into the infrared.

Gathering the multispectral data is only the first step in recovering the erased (or damaged) text. The goal is to record the differences in the reflectance or fluorescence of the different inks and the parchment under different wavelengths of illumination. If these differences are easily visible at one or more of the wavelengths, then the text can be easily separated. On the other hand, these differences might manifest themselves in slight, low-contrast variations spread across the wavelengths, and not readily visible to the eye at any one wavelength. This is where sophisticated methods of extracting low contrast differences, through the image processing of the multispectral data, are needed to make the erased writing visible for the scholar to read.

Simple image processing methods that utilize the visible differences in wavelength response between the visible and the red, or near infrared, exposures were developed for the Archimedes Palimpsest Project and expanded for the Sinai Palimpsests Project. These methods are implemented in custom software, written for UNIX in C, and are the subject of this paper.

On the Sinai Palimpsests Project, other methods have been explored that use algorithms such as Principal Components Analysis (PCA) and Independent Components Analysis (ICA). These algorithms are capable of extracting image differences that are not readily visible to the eye, making them a subsequent step in processing, when simple processing methods do not yield sufficient results.

It is an art to use these sophisticated algorithms and the other image processing team members of the Sinai Palimpsests Project have spent several years honing their skills in this area. These other scientists are Dr Roger Easton, Jr of the Rochester Institute of Technology, Dr David Kelbe of the Oak Ridge National Laboratory and Dr William Christens-Barry of Equipoise Imaging LLC. The use of these sophisticated algorithms is not within the scope of this paper.

Fig. 4: A pseudocolor image is formed by combining two images into a single color image.

3. Sinai Palimpsests Project

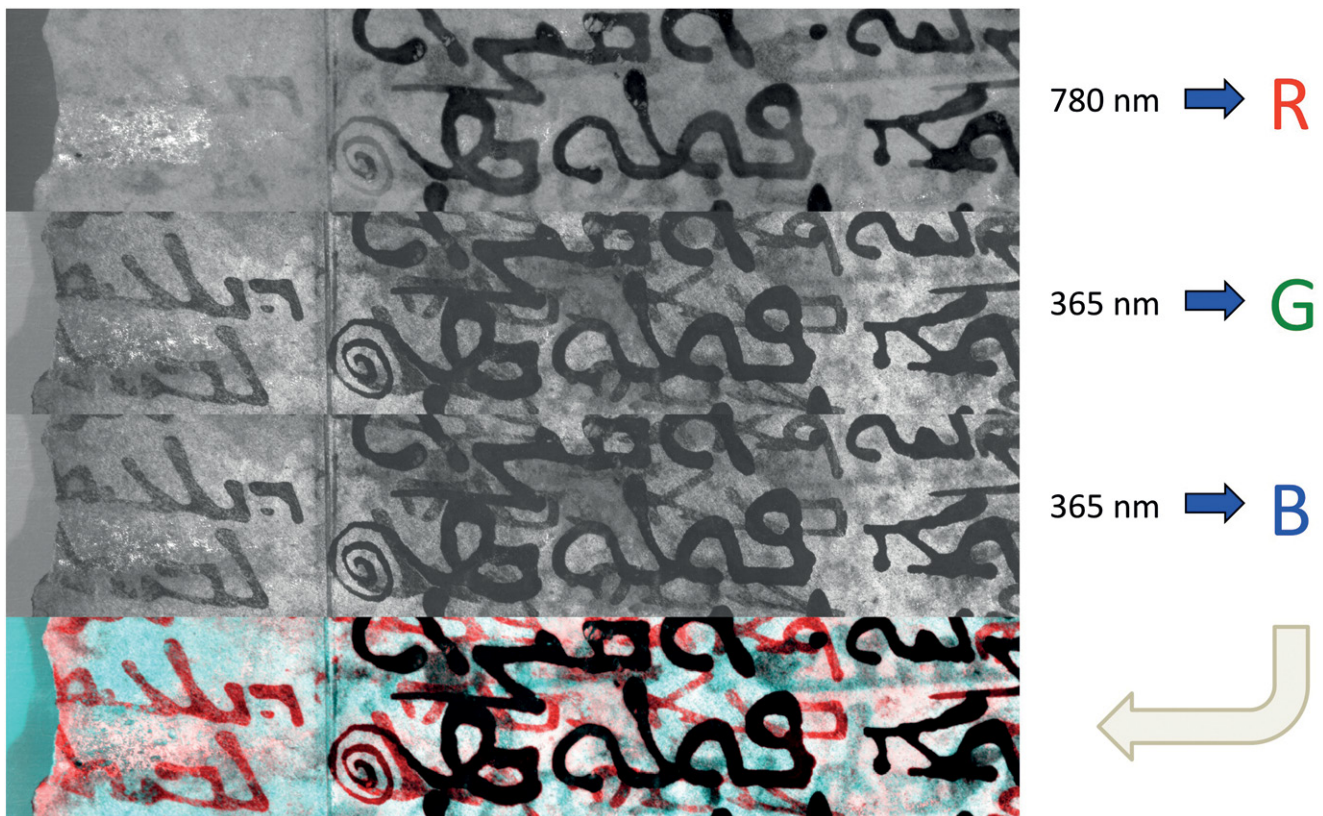The Sinai Palimpsests Project is a 5-year project that started in 2011. It is a joint effort of the Holy Monastery of St Catherine at Mount Sinai, the Early Manuscripts Electronic Library of California, and the Arcadia Foundation of the United Kingdom, to image palimpsests within the monastery's library and make them available to scholars.[3]

There are approximately 160 known palimpsests in the library of the Holy Monastery of St Catherine. Around 72 of these palimpsests have been imaged and processed by the project team. All images have been processed with the standard processing methods that will be described here.

The primary method is to form a pseudocolor image from two wavelengths of the multispectral image set.[4] The primary wavelengths that are combined are an ultraviolet image and a visible image. The ultraviolet image may be the 365 nm image shown in the upper left corner of Fig. 3, or it may be one of the filtered ultraviolet images, not shown in that figure. The second image is a visible image that is typically taken from the red to infrared region, perhaps the 625 nm image or the 780 nm image.

These two images are combined into a single RGB color image, as shown in Fig.4. The visible image, in this case at 780 nm, shows the overwriting, but the erased writing has almost completely disappeared into the parchment background. On the ultraviolet image, the erased writing stands out, along with the overwriting. By this method, the erased writing appears bright in the red separation and dark in the green and blue separations, making it appear to be red in the pseudocolor image. On the other hand, the overwriting is dark in all three color separations, making it appear neutral gray or black in the pseudocolor image. As a result, the erased writing shows up in high color contrast with respect to both the overwriting and the parchment background. This makes it easy for the scholar to read, even though some of the letters are partially obscured by the overwriting.

One additional step is done to each of the two images (the ultraviolet image and the visible image) before the two images are combined. That step is a locally-adaptive, normalization of the contrast of the image detail. A sliding window is moved across the image. Within the window, the

---

[3] *Sinai Palimpsests Project* <http://sinaipalimpsests.org>.

[4] Knox 2008.

Fig. 5: The natural light image of Holy Monastery of St Catherine at Mount Sinai, Ms. Syriac 30, fol. 4ʳ on the left, compared to the pseudocolor image on the right.

Fig. 6: Natural light, Holy Monastery of St Catherine at Mount Sinai, Ms.Georgian NF 13, fol. 59ʳ.

mean and the variance are measured. With these statistics, the center pixel of the sliding window is put through a linear stretch that adjusts the mean value to 127.5 and stretches the contrast so that ±3 standard deviations extend linearly between black and white (0–255).

This is a local adjustment, because the amount of contrast stretching depends on the image values within the sliding window. This tends to smooth out the variations in the background levels across the leaf and to equalize the contrast of the characters against the parchment background. That makes it easier to make the overwriting a neutral color, while enhancing the color contrast of the erased writing. The final result of the pseudocolor process can be seen for the whole leaf in Fig. 5.

A second standard processing method that was applied to all of the images on the Sinai Palimpsests Project is called a transmission ratio image. Images captured in transmission were taken of each leaf with a custom sheet illuminator on which each leaf is placed. Light from LEDs is fed into the clear transparent illuminator and it radiates uniformly out of the sheet, through the leaf and into the camera above. This

enables the camera to see light, in wavelengths through the visible and into the infrared, that travels through the leaf from the back of the parchment. The reflectance images, as described earlier, are taken along with the transmission images. Since the leaf does not move during either process, the transmission images are registered with all of the other images.

The transmission ratio image is formed by taking the ratio of the transmission image at 940 nm by the reflectance image at 940 nm. Since the overwriting shows up in both images, this ratio tends to reduce the contrast of the overwriting. A contrast enhancement and a localized sharpening are then performed. The importance of the transmission ratio image can be seen in the next example. In Fig. 6 is shown a natural light image of Georgian NF 13, fol. 59 recto. The erased writing is very difficult to see. There appear to be some horizontal smudges where the erased characters should be. On the right-hand side of the image, in the center of the leaf, there are some characters that are barely visible.

The pseudocolor image of this page is shown in Fig. 7. There are some characters visible at the top of the image in
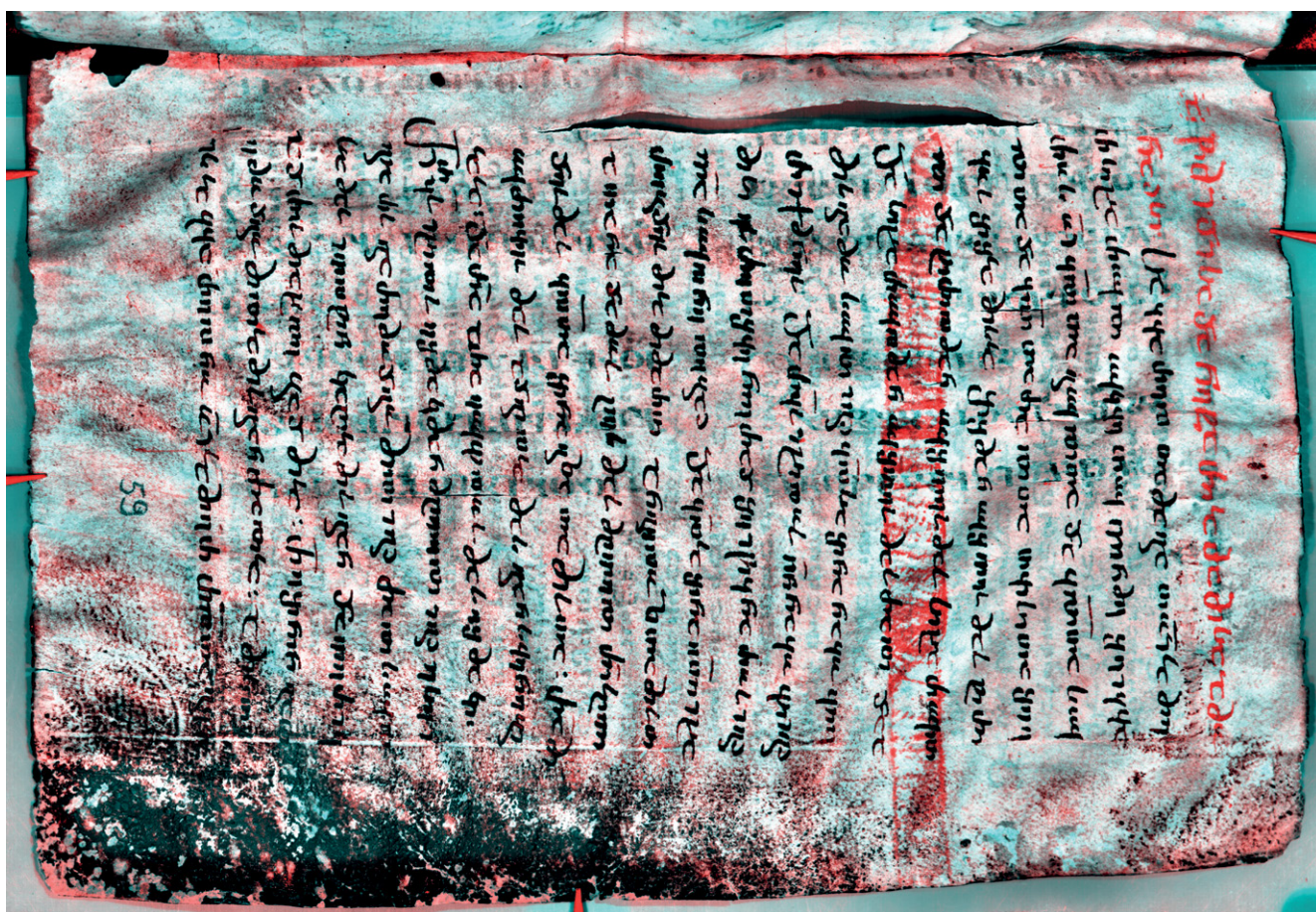

Fig. 7: Pseudocolor, Holy Monastery of St Catherine at Mount Sinai, Ms. Georgian NF 13, fol. 59ʳ.
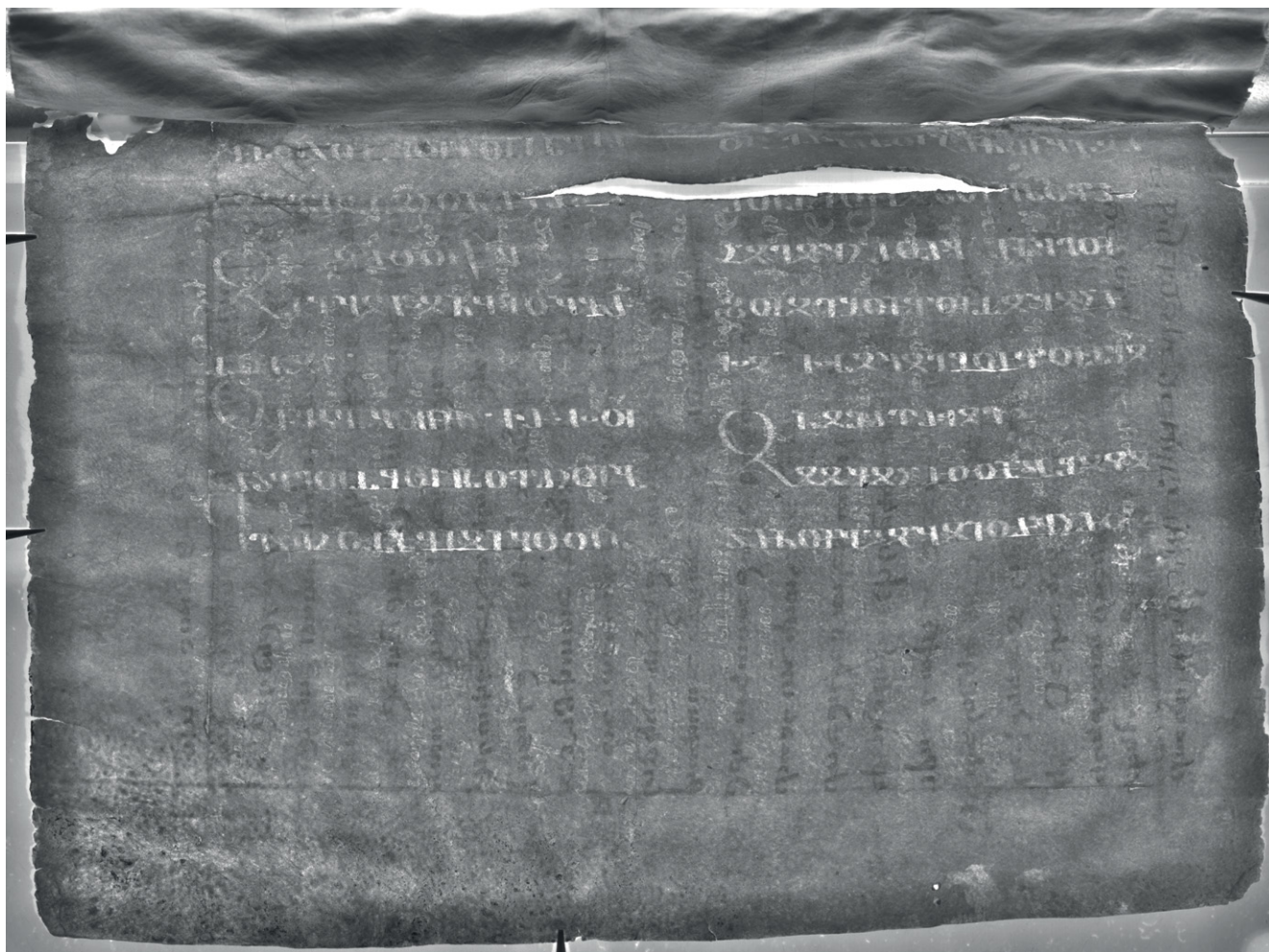
Fig. 8: Transmission ratio, Holy Monastery of St Catherine at Mount Sinai, Ms. Georgian NF 13, fol. 59ʳ.

the gutter. For the rest of the leaf, though, there are hints of cyan colored characters, but nothing like the red characters typical of a fluorescence image taken under ultraviolet illumination.

The problem is that the residual acid of the original ink has eaten into the parchment. The parchment is a little thinner in the region of the character and no stain exists there to inhibit the fluorescence. As a result, the region of the character fluoresces along with the parchment, resulting in very low contrast characters in the pseudocolor image.

The fact that the parchment is thinner, though, means that the transmission image is now very valuable. In the transmission image, more light makes it through the leaf specifically in the regions where the erased characters have eaten into the parchment. This produces an image that shows clear characters that are otherwise very difficult to see. Since this process of eating away at the parchment tends to happen preferentially on the flesh side of the parchment, only the characters from one side are visible in the transmission ratio image.

The transmission ratio image of Georgian NF 13, fol. 59 recto is shown in Fig. 8. Now the characters that were barely visible in some regions, and not visible at all in most regions, are now clearly visible in transmission. Although this is a very simple process, the transmission ratio image has turned out to be a very valuable addition to the image processing toolkit for the Sinai Palimpsests Project.

4. Image Processing Software

The software to produce the pseudocolor and the transmission ratio images in this paper is part of a UNIX-based package of image processing routines, written in C by the author, between 2000 and 2013, to process the multispectral images of the Archimedes Palimpsest project.[5]

The UNIX operating system has the advantage that processing modules can be created (each implementing a

---

[5] *The Archimedes Palimpsest* <http://www.archimedespalimpsest.org>.

```
●○○                    Terminal — tcsh — 87×8
----pali:/Users/knox----
<1> cd /volumes/Knox_Chartres/Hamburg/Syriac30/Flattened_Images/
----pali:/volumes/Knox_Chartres/Hamburg/Syriac30/Flattened_Images----
<2> readtif K0047_000007+MB365UV_pack16.tif | div -f 0047_000007+MB940IR_pack16.tif |
packimage -s | rotate -a -90 | show &
[1] 4821 4822 4823 4824 4825
----pali:/volumes/Knox_Chartres/Hamburg/Syriac30/Flattened_Images----
<3> ▯
```

Fig. 9: Text-based interface of the UNIX-based, C-language image processing system. In this example, a ratio of two images is displayed.

single algorithm), that run independently and simultaneously, and communicate with each other by exchanging image scanlines over UNIX pipes. The use of UNIX pipes to exchange image scanlines, means that only a few scanlines of the image are in memory at any one moment. Each module receives a scanline, processes it and passes it onto the next module, before retrieving its next scanline. In this way, a very large image can be easily processed without running out of memory or requiring a computer with large memory stores.

Since each algorithm is implemented in its own module, standard algorithms can be implemented and tested separately, and then used without change at a later time. To add a new algorithm, a new module is written and tested without introducing errors in the implementations of the existing algorithms. To use the new module, it is simply included in the UNIX command line. This is easy for a software researcher to do, but is beyond the capabilities of a non-technical user.

For example, a command line to take the ratio of two images, using the UNIX-based software, is shown in Fig. 9. A file is read and piped to a routine that reads a second image and takes the ratio. This result is piped to a routine that adjusts the contrast and lastly to a routine that rotates the image for display. For the expert, this is easy to set up, but tedious to type into a text window. For the non-expert, this tends to be too complicated to be useful.

To make the software available to more people, a new version of the software, called Hoku, is being created in the Java language. The move to the Java programming language was made for two reasons. First, Java is a portable language that is available on almost all computers and operating systems. Secondly, Java comes with tools that make it easy to create graphical user interfaces. These two features make it possible to create an image processing package that can be used by a large number of people with varying degrees of technical expertise.

Although Java does not implement UNIX pipes, within Hoku, a modular architecture was created to enable each
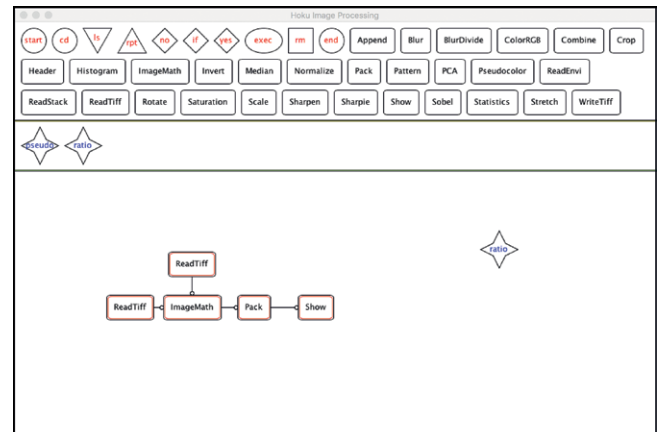

Fig. 10: User interface for the Java-language based Hoku image processing system.

module to be run as an independent software 'thread' with an interface that enables modules to retrieve and send processed scanlines. As a result, a new image processing capability can be easily incorporated into the package. Since each module runs as an independent 'thread', different modules can run on different processors of a multi-processor computer, shortening execution times.

The Java software package is still under development, but a preliminary user interface is shown in Fig. 10. The interface consists of three regions, a 'cupboard' on the top, a 'shelf' in the middle and a 'desktop' on the bottom. A list of available routines is automatically created as the package starts up and is displayed along the top in the 'cupboard'. To use one of the modules, the user simply drags it into the 'desktop' with the mouse or track pad. As multiple modules are added to the processing task, links are automatically connected between modules. A small circle on the link acts like an arrow and indicates the direction of flow of the image scanlines. The non-rectangular modules at the top left, with names in red, are control modules, which will enable jobs to be run on complete directories of images.

In the example shown below, two images are read, fed to a module called ImageMath, where a ratio is taken. The ratio is contrast adjusted and displayed on the screen. This job can be run in batch mode. If one draws a box around the collection of modules, that collection is collapsed into a 'star' figure containing the name of the job, in blue letters. This job can be dragged to the 'shelf', where it can be retrieved at a later time. When dragged to the 'shelf', a command file containing a source description of the job is saved to disk for later use. The job can be re-opened on the 'desktop' by drawing a box around the job icon. This

expands the job into its set of linked modules allowing additional editing.

On the 'desktop', a job can be executed by right-clicking within the job icon and selecting 'Execute' from the pop-up menu that appears. When this task is executed, each ReadTiff module opens its image file, reads and feeds individual scanlines to the next module down the pipeline. Every other module, all of them running independently and simultaneously, gets a new scanline (from the module before it), processes the scanline, and then passes the processed scanline to the next module in the pipeline. In this way, images of arbitrary sizes can be processed with this software, without requiring large amounts of computer memory. Any module can buffer a few input scanlines as needed to produce an output scanline. Multiple jobs can be created, edited and executed on the 'desktop' at the same time.

There are commercial image processing systems available to process multispectral imagery. For example, ENVI is a software package developed for remote sensing image data.[6] While these commercial packages contain many image processing features, typically, they are complicated and the cost of the software can put it out of the reach of many potential users.

An open source Java software package, called ImageJ, also exists.[7] This software is freely available and has many algorithms available through plug-ins. It would be possible to implement the image processing software described in this paper with ImageJ, but it would lack one main advantage, that of small memory. The ImageJ software reads the whole image into memory, requiring a computer with a lot of memory to hold large images. The Hoku software package will have the advantage of being able to operate on computers with a small amount of memory, but still process large images.

In comparing the amount of memory required, ENVI typically will create an image cube that contains all of the images for a given leaf, stored in 64-bit pixels, or floating-point doubles. The raw images from MegaVision are currently 100 MB each, i.e. 16 bits/pixel and 8176 × 6132 pixels. Currently, 50 images are taken for each leaf. That means that one image cube requires 20 GB of image memory

just to hold one copy of it in ENVI's memory. ImageJ typically reads the 16-bit pixels, but that still requires 5 GB for one image cube. Any processing within ImageJ, might require additional copies of that image cube. On the other hand, Hoku holds one or maybe a few scanlines at a time in memory in each module. For example, a job with 10 modules, where each module has an input buffer, a working buffer and an output buffer, would require 30, let's say 100 scanlines. If each of those 100 scanlines were floating point, Hoku would require 300 MB of memory to process the 20 GB image cube.

Today, computers with large memories, such as 32 GB, are becoming more readily available. Such a computer could hold one 20 GB image cube in memory. Why not write the software to hold everything in memory to take advantage of the new computer capability? One reason is that not everyone can afford to purchase such computers. Secondly, when the prices of computers eventually do come down to affordable levels, new camera technology will also be available. That new camera technology will have increased resolution and will produce captured image data of increased size. As a result, once again the amount of memory needed to process the new image cubes will exceed the capability of affordable computer memory. It is far better to have an image processing system for which the size of the computer memory does not limit the size of the image that can be processed.

Hoku, the Java package described in this paper, will be distributed free of charge. Initially, it will not contain much of the sophisticated capability of the ENVI system, or of ImageJ, but it will provide the capability described in this paper. Currently, only the author is developing this software package. By the end of 2018, the package will be sufficiently developed to enable additional developers to join the effort. At that time, the Hoku Java jar file and Hoku sources will be released on GitHub. As other people add modules, the available capability will grow. In addition, because the system is versatile and easily adaptable, the software package can be tailored to any image processing requirements, not just the software described here.

---

[6] Harris Geospatial Solutions, *Linear Spectral Unmixing* <http://www.harrisgeospatial.com/docs/LinearSpectralUnmixing.html>, <https://www.exelisvis.com/docs/linearspectralunmixing.html>.

[7] *Wikipedia*, 'ImageJ' <http://en.wikipedia.org/wiki/ImageJ>.

REFERENCES

Easton, Jr., R. L., K. T. Knox, W. A. Christens-Barry, K. Boydston, M. B. Toth, D. Emery, and W. Noel (2010), 'Standardized System for Multispectral Imaging of Palimpsests', *Proceedings of SPIE 7531, Computer Vision and Image Analysis of Art, San Jose, California*. <doi:10.1117/12.766679>.

Harris Geospatial Solutions, *Linear Spectral Unmixing* <http://www.harrisgeospatial.com/docs/LinearSpectralUnmixing.html>, <https://www.exelisvis.com/docs/linearspectralunmixing.html> (last accessed 15 April 2018).

Knox, K. T. (2008, February), 'Enhancement of Overwritten Text in the Archimedes Palimpsest', *Electronic Imaging 2008* (International Society for Optics and Photonics), <doi:10.1117/12.766679>, 681004–681004.

MegaVision Archival and Cultural Heritage Imaging <http://www.mega-vision.com/cultural_heritage.html> (last accessed 15 April 2018).

Sinai Palimpsests Project <http://sinaipalimpsests.org> (last accessed 15 April 2018).

The Archimedes Palimpsest <http://www.archimedespalimpsest.org> (last accessed 15 April 2018).

Wikipedia, 'ImageJ' <http://en.wikipedia.org/wiki/ImageJ> (last accessed 15 April 2018).

PICTURE CREDITS

Figs 1–3: © The author.

Figs 4–8: © Holy Monastery of St Catherine at Mount Sinai.

Figs 9 and 10: © The author.

Studies in Manuscript Cultures (SMC)

Ed. by Michael Friedrich, Harunaga Isaacson, and Jörg B. Quenzer

Writing is one of the most important cultural techniques, and writing has been handwriting throughout the greater part of human history, in some places even until very recently. Manuscripts are usually studied primarily for their contents, that is, for the texts, images and notation they carry, but they are also unique artefacts, the study of which can reveal how they were produced and used. The social and cultural history of manuscripts allows for 'grounding' the history of human knowledge and knowledge practices in material evidence in ways largely unexplored by traditional scholarship.

With very few exceptions, the history of the handwritten book is usually taken to be the prehistory of the (printed Western) book, thus not only denying manuscripts their distinct status as carrier medium, but also neglecting the rich heritage of Asian and African manuscript cultures from which, according to conservative estimates, more than ten million specimens survive until today.

The series *Studies in Manuscript Cultures (SMC)* is designed to publish monographs and collective volumes contributing to the emerging field of manuscript studies (or manuscriptology) including disciplines such as philology, palaeography, codicology, art history, and material analysis. SMC encourages comparative study and contributes to a historical and systematic survey of manuscript cultures.
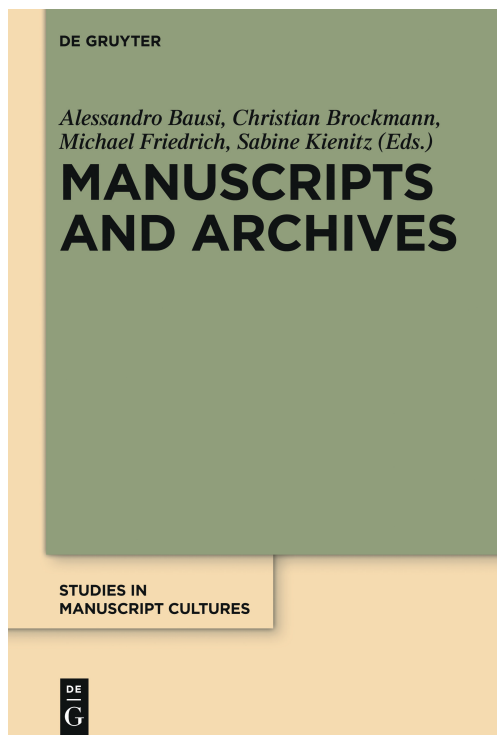
Just published

DE GRUYTER

Alessandro Bausi, Christian Brockmann,
Michael Friedrich, Sabine Kienitz (Eds.)

MANUSCRIPTS AND ARCHIVES

STUDIES IN
MANUSCRIPT CULTURES

DE G

11 – *Manuscripts and Archives: Comparative Views on Record-Keeping* edited by Alessandro Bausi, Christian Brockmann, Michael Friedrich, and Sabine Kienitz

Archives are considered to be collections of administrative, legal, commercial and other records or the actual place where they are located. They have become ubiquitous in the modern world, but emerged not much later than the invention of writing. Following Foucault, who first used the word archive in a metaphorical sense as 'the general system of the formation and transformation of statements' in his 'Archaeology of Knowledge' (1969), postmodern theorists have tried to exploit the potential of this concept and initiated the 'archival turn'. In recent years, however, archives have attracted the attention of anthropologists and historians of different denominations regarding them as historical objects and 'grounding' them again in real institutions. The papers in this volume explore the complex topic of the archive in a historical, systematic and comparative context and view it in the broader context of manuscript cultures by addressing questions like how, by whom and for which purpose were archival records produced, and if they differ from literary manuscripts regarding materials, formats, and producers (scribes).

www.manuscript-cultures.uni-hamburg.de