https://doi.org/10.2352/ISSN.2470-1173.2019.14.COLOR-100

Creating a simulation option for the reconstruction of ancient documents: Palimpsests

Reiner Eschbach[1], Roger L. Easton[2], Keith T. Knox[3], Jon Y. Hardeberg[1]

1) Norwegian Technical University, NTNU Dept. Computer Science, Teknologivegen 22, Gjøvik, Norway
2) Rochester Institute of Technology, RIT, Chester F. Carlson Center for Imaging Science, Rochester NY, 14623 USA
3) EMEL Early Manuscripts Electronic Library, 904 Silver Spur Road, #254, Rolling Hills Estates, California 90274

Abstract

Ancient documents were often created by overwriting older documents that had been erased to re-use the expensive parchment. There is thus the situation that we have valuable documents that actually contain possibly more valuable documents "underneath", hidden to the human eye. Being able to retrieve these originals is an important task, however, associated with several problems. One of these is that the value of the documents limits the distribution and thus the ability to experiment. Another is that the "to be discovered" underwriting might hold important information that should first be examined, classified and put into context by textual scholars, again severely limiting the ability to experiment. This is a strong decrement to scientists and students involved in developing and testing imaging algorithms.

This paper describes an approach to create artificial palimpsests that are reasonable approximations based on previous example and thus can be used by everybody to test new assumptions, new algorithms and to study the interaction of the different deterioration mechanisms.

Introduction

Ancient documents are valuable for a variety of reasons, one of them being the possibility to discover, rediscover or learn things about the geographical, political, societal and technological situation at the time of the writing of the document. But there is an even more intriguing side to many ancient documents. The writing medium – often parchment – was rare and expensive and thus many of these ancient texts are written on a medium that previously held an even older text. Often these older texts were erased, washed off, or scraped off, and the parchment was then used as the medium for a newer text, generally even changing formats, sizes, bindings, etc. These overwritten documents (or erased documents) are called Palimpsets. The erased part is normally not visible to the unaided eye and in the past decades a very successful effort was made to utilize multispectral imaging to recover the underwriting. However, the main focus was always on the immediate document and thus a multitude of approaches, many highly manual have been developed with very little effort to understand and examine underlying imaging models.

Figure 1 shows an example of such Palimpsest. In this case, the underwriting was not well erased and even to an unaided eye it is clear that there is an original text hidden and – hopefully – recoverable. As a rule of thumb one can say that the data shown in Figure 1[1] is an extremely easy case and that most documents of this type have been deciphered.

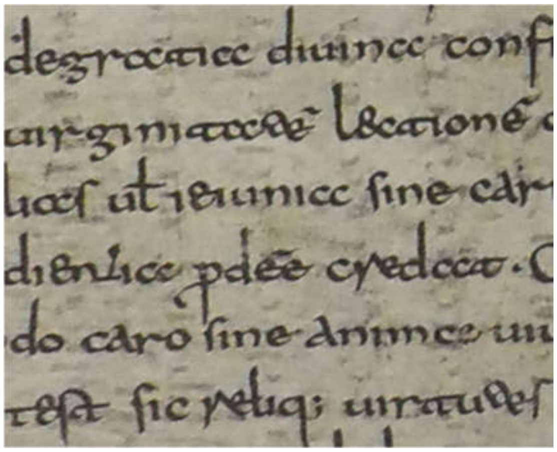


Figure 1. Example of a simple Palimpsests (taken from (1))

Motivation

In order to test new approaches, new algorithms and new ideas, it is important to have a representative test-set of images available. With historically valuable documents, this poses a problem. "Document Ownership" is normally not an issue for research in imaging. In historical documents, however, there is the physical value of the artifact, meaning that scans are done by view groups under very controlled settings, and there is also the historical value of the content, meaning that images of recovered text should not be shown or published until they have been examined and classified by textual scholars. This means that only very few actual Palimpsest data exist that are freely available to try new ideas.

[1] Taken from "Spicilegium palimpsestorum arte photographica paratum", Harrassowity 1913 (Google Books)

Additionally, the existing Palimpsests do not represent a good test set to try new approaches. As an example: if current recovery algorithms are sensitive to misregistration, no recovered Palimpsest will exist that shows misregistration. Thus a new algorithm has no ground-truth to compare against.

The described simulation environment is intended as an approach to reduce this problem by allowing believable Palimpsest simulations that can be varied in a controlled manner.

Simulation Environment

For the simulation environment we are trying to build this data serves as a starting point with all the simulated documents being of "higher difficulty", where it is understood that "difficulty" in this sense reflects the opinions of the authors and not an objective metric.

In the following, we will describe the individual parameters we will use for the simulation environment, and we will show how the simulated Palimpsests respond to the standard recovery algorithms used.

Parameter Space

In order to simulate a palimpsest we have to define the parameters that are likely to be contributing and/or under the control of the experiment. In our scenario we started from

- Number of Separations
- Dynamic range
- Noise
- Number of Signal separation
- Misregistration (various types)
- Physical deterioration

And we will describe them in more detail below.

1) Number of Separations

In the current simulation environment, the number of separations is a free parameter and only limited by the compute time. The current measurement system employed by some of us[2]allows up to 42 separations, and one of the important question that will be asked from the simulation environment is the trade-off between time, number of separations and other parameters.

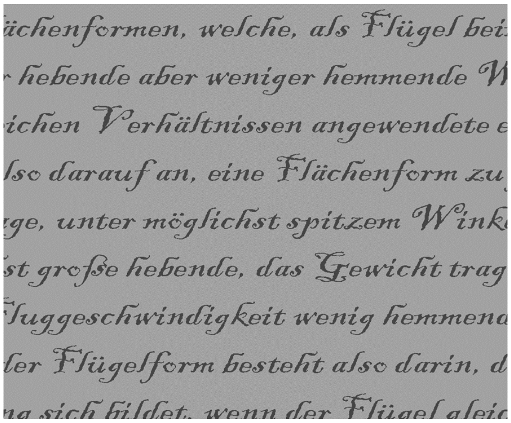
For the purpose of the simulation inside this Proceedings Paper, we will use 15 separations as a compromise between the 3 channel rgb and the multi-spectral 42 bands.

2) Dynamic Range

The Dynamic Range parameter is simply describing at what bit-depth the underwriting exists compared to the overwriting. In Figure 1, the underwriting was noticeable in an 8-bit image and thus we decided to have all our simulations at a lower range. From an existing Palimpsest that had previously be recovered, we set the ranges as follows:

a) White of paper ~ 25000
b) Black of Overwriting ~ 7000
c) Range of erased underwriting ~20

This would – in a noise free simulation- result in a paper of 25000 or 24980 and in an overwriting of 7000 or 6980, meaning we are using an additive imaging model for the underwriting. Of course, the actual levels are independently set for all separations. Figure 2 shows such an example. Even in this unrealistic case of "zero noise", the underlying data is no longer visible to the human eye (and also not in the image of Figure 2)



ächenformen, welche, als Flügel bei
r hebende aber weniger hemmende W
ichen Verhältnissen angewendete e
lso darauf an, eine Flächenform zu
ge, unter möglichst spitzem Wink
st große hebende, das Gewicht trag
Fluggeschwindigkeit wenig hemmend
ler Flügelform besteht also darin, d
na sich bildet, wenn der Flügel gleic

Figure 2. Noise-free simulation with an underwriting at approx. 1/1000 of signal.

Noise

In order to estimate reasonable noise levels, we simulate palimpsests with the above data at different noise levels and ran them through standard Palimpsest recovery algorithm. In our case we decided to use a PCS method since the performance between PCA and ICA was comparable and PCA is more widely used. Figure 3 shows the result for the noise levels of σ="70", "110" and "150".

Magnifying the data (and rotating by 90°) we get Figure 4, with the top row representing a noise level of "70", the middle "100" and the bottom "150".

[2] R. Easton & K. Knox

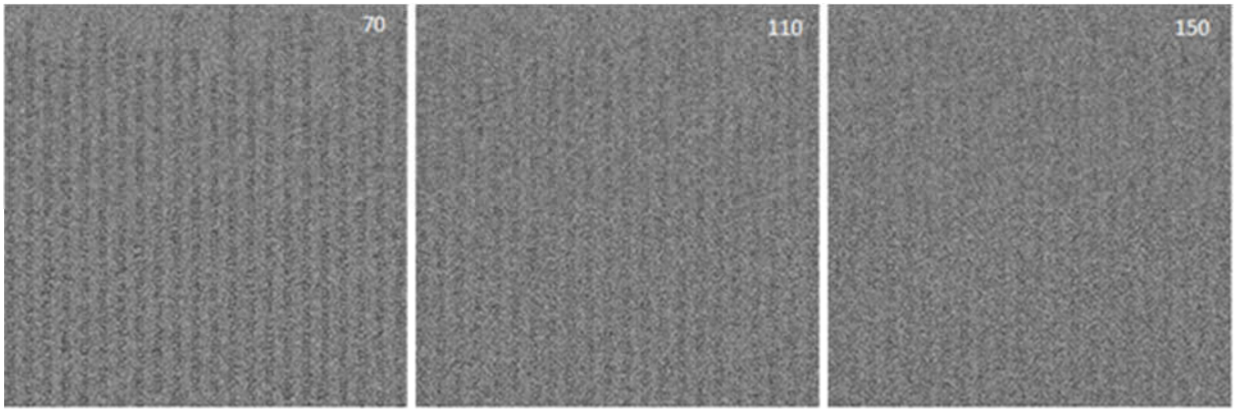


Figure 3. Single separation reconstruction with noise level as indicated in the top right corner.

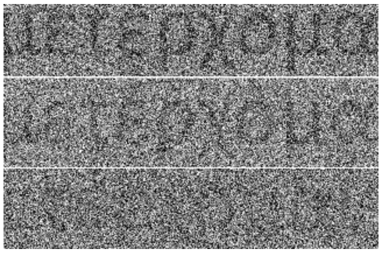


Figure 4. Magnification of a crop area from Figure 3 (rotated).

This simulation showed us that we can simulate cases from “unreadable” to “easily readable”. If one wants to explore the effect of other deteriorations than noise we can thus create “low noise” scenarios, as well as “at the edge” scenarios where the noise almost makes the data unreadable.

Since we are interested in exploring the effect of the other parameters, we will use a noise level of 100 in our subsequent simulations. Note that this means that the 1□ of our noise is about 5 times the level of the underwriting.

The noise level can independently set for all separations.

3) Number of signal separations

When scanning the original document in n-separations, it is not clear if all separations actually contain signal from the underwriting. Adding separations that do not contain signal might impact the overall reconstruction and in our scenario we can determine the subset that contains signal by simply modifying the signal dynamic range through item (2).

4) Misregistration

When scanning n-separations, the individual separations might be misregistered via several mechanisms.

5.1) Vibration

One mechanism for misregistration is any vibration between separations which results in an image shift. Information gained from this experiment will give data for the system requirements. In Figure 5, the following parameters were used:

(a) No misregistration: reconstruction in the first PCA band
(b) A random single pixel shift with 40% likelihood of the separations: reconstruction in PCA band #7

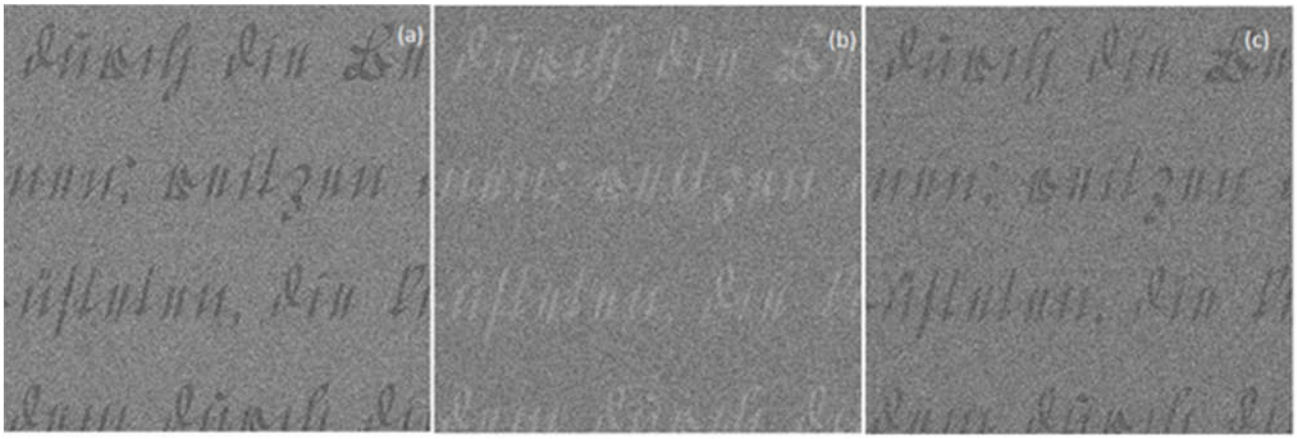


Figure 5. Three reconstructions of misregistration examples, where the image represents the "best" PCA band.

(c) A random single pixel shift with 43% and a random double pixel shift with 13% likelihood of the separations: reconstruction in PCA band #9

Figure 6. Three misregistration examples where the images reflect different PCA bands dependent on the misregistration.

The reconstruction in higher PCA bands is a very common artifact of many palimpsests and the simulations can also exhibit this effect – albeit it is not proven that the mechanism is identical.

5.2) Chromatic aberration

Additionally, the system can simulate a chromatic aberration where the separations are scaled with respect to each other. The "optical axis", i.e.: the center for the aberration can be freely selected and the axis is not restricted to the center of the image. Additionally, the insertion of a filter can be simulated by having multiple different scalings for different separation bands.

6) Physical deterioration

Ancient documents have undergone physical deterioration, be it stains, rub-offs, dirt, or any other of a number of processes. We simulate these artifacts by using actual photos ore renderings of physically damaged paper.

Figure 6 shows the flow of the entire simulation system.

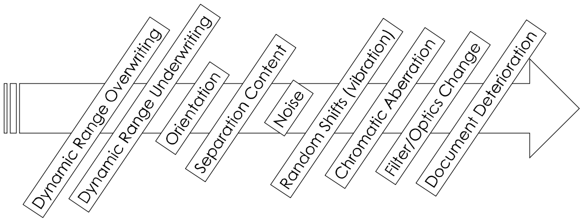


Figure 6. Flowchart of the simulation system.

Initial Results

In order to test if the simulated images are "believable", we used them as starting points for existing recovery scripts and compared them – visually – to results obtained from real data sets. For our tests, we – of course – new the used distortions, knew that the underwriting is at 90° to the overwriting, etc. These tests were not intended to examine new algorithms, but simply to get an estimate of the critical parameters.

Figures 7 and 8 show two instances. In Figure 7, we chose a physical deterioration that was relatively homogeneous across the page and in Figure 8 we chose one that has a high contrast edge, attempting to see the sensitivity of the reconstruction of the spatial content of the overall image. The right sides show the reconstructed test (picking the visually best PCA band).

From a pure visual inspection, it seems that the simulation data is a believable reflection of actual Palimpsests data. The left side reflects the look of a source document and the reconstruction on the right side has similar characteristics to other reconstructions of actual data. The simulation thus should be a good source to test algorithm sensitivity, e.g.: with respect to misregistration.

Summary

We are trying to build a simulation environment that can be used to simulate believable Palimpsests where certain artifacts or deterioration can be precisely modelled and varied. This, as we hope, would allow us to test registration sensitivity – at what misregistration does the algorithm start to fail - , allow us to test the compromise for separation numbers – when do additional separations without signal deteriorate the results-, allow us to test filter settings – how does the best filter size depend on the overwriting structure -, etc.

It is intended that the simulated Palimpsests will be made publically available with a set of defined distortions and a set of "blind" data where the actual deterioration is hidden from the experimenter.

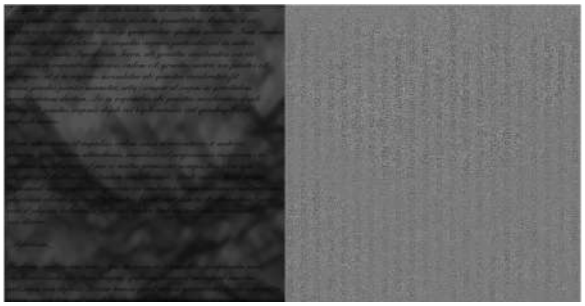


Figure 7. A simulated Palimpsest and the reconstruction of the underlying text.

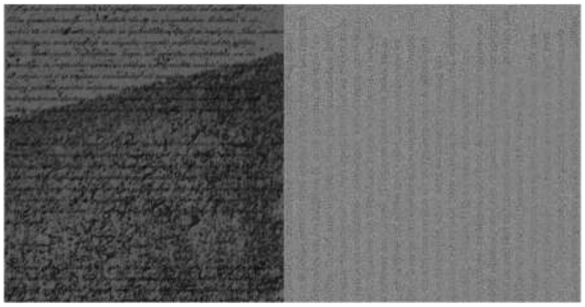


Figure 8. Another simulated Palimpsest and its reconstruction.

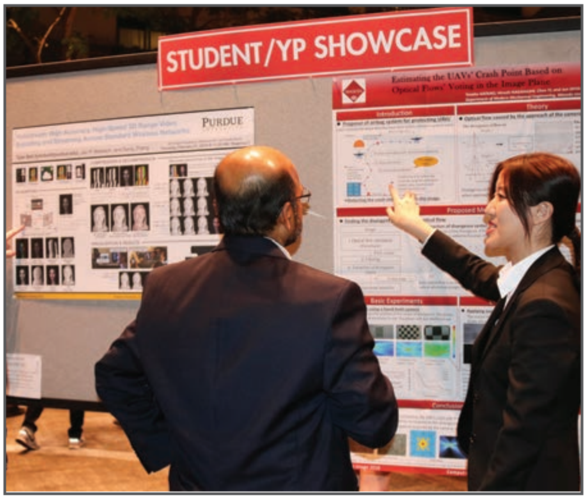
JOIN US AT THE NEXT EI!

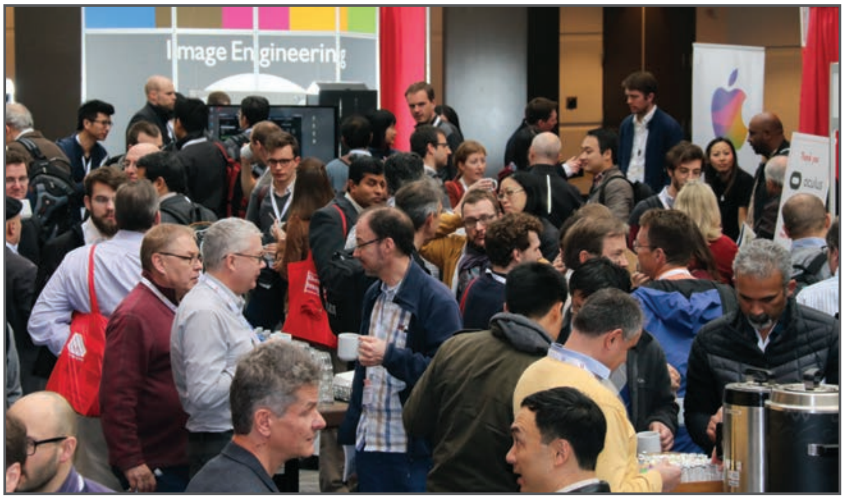
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org