TEACHING HANDWRITTEN DOCUMENT RESTORATION TECHNIQUES USING STATISTICAL PROCESSING

Osvaldo Clúa [1] *and María Feldgen* [2]

Abstract *— We digitalized a corpus of handwritten documents from an Heritage Preservation Program. Several documents were severely damaged due to time and storage environment. In a traditional image processing course, several filters will be applied to recover the images from selected degradations. To understand these filters, students must have had an exposure to signal analysis techniques. We departed from the traditional scheme using statistical image analysis techniques. Students were able to perform some measurement of the degradations and their recovery without restoring to signal analysis concepts. The degraded image is statistically modeled and the observation probability density is obtained from the gray-level histogram of the input image. The use of historical documents motivated our students as they were able to recognize in the document historical situations they have studied in the high school*

Index Terms — Image Processing, Introduction Courses, Classroom Assessment Techniques, Math Awareness

INTRODUCTION

In image processing fundamental books it is common to find expressions like

$$\frac{ds}{d\tau} = (L-1)\frac{d\left[\int_0^{\tau} p_r(w)\,dw\right]}{d\tau}$$

(Gonzalez-Woods 2008 p 124) which in several cases boils up to a simple $\frac{1}{L-1}$

Considerable mathematics skills are required for research or advanced work in Image Processing. In order to enhance or process some image features, concepts from signal analysis are needed and developed as images are considered two dimensional signals. Signal analysis is not a mandatory subject, and is regarded as difficult by a great number of informatic's students. Advanced techniques such as feature extraction or Region of Interest Management are introduced after a long mathematical prelude.

An image may be considered as an array of various dimensions. The processing of this array often requires considerable mathematical skills. However, almost all of the introductory material can be taught using elementary math and statistic processing.

Understanding an Image Processing topic does not mean understanding all the theoretical underpinnings of it. An appreciation of the effect of an algorithm on the image rendering can be developed without the overhead of difficult mathematics. These approach has been used for example in[1].

Image processing can also be infused into existing curricula, in courses such as CS1 or CS2 to motivate students with some applied approach, as reported in [2]. In the referenced work authors used the image processing Java classes to introduce topics such as array processing and tree traversal. A similar approach using C in CS1 is found in [3].

THE PROGRAM

The Computing Curricula includes Image Processing only in their Computer Engineering course recommendation [4], in a Signal and System Context:

Digital Signal Processing (CE-DSP)

1. CE-DSP0 History and overview [core]
2. CE-DSP1 Theories and concepts [core]
3. CE-DSP2 Digital spectra analysis [core]
4. CE-DSP3 Discrete Fourier transform [core]
5. CE-DSP4 Sampling [core]
6. CE-DSP5 Transforms [core]
7. CE-DSP6 Digital filters [core]
8. CE-DSP7 Discrete time signals [elective]
9. CE-DSP8 Window functions [elective]
10. CE-DSP9 Convolution [elective]
11. CE-DSP10 Audio processing [elective]
12. CE-DSP11 Image processing [elective]

As shown, Image processing is not in the core recommendation but an elective recommendation after developing the necessary background.

This background on Signal Processing is not present in the Computer Science Curriculum [5]. However, Image processing can be an engaging topic for students.

We decided to introduce some concepts on Image processing at a sophomore level, as an elective subject using elemental math.

The topics covered are:

1. Definition of a digital image; digital images as matrices, types of images (binary, greyscale, colored, indexed).
2. Image file formats: some examples and definitions. Use of different formats for programming.

[1] Osvaldo Clúa, Univesidad de Buenos Aires, oclua@ieee.rog
[2] María Feldgen, Universidad de Buenos Aires, mfeldgen@ieee.org

3. Point operations: thresholding and histogram processing.
4. Basic topology: neighbors and neighborhoods, adjacency and connectedness, boundary following and object counting.
5. Spatial filtering: low and high pass filters. 5. Edge detection and enhancement. 6. Noise: types of noise (salt & pepper, gaussian, speckle, periodic). Removal methods: average filtering, rank-order filtering, Wiener filters.

As a programming support we studied Matlab [6] (proprietary) or its GPL brother Octave [7] and the free R Project software [8]. Both systems are used subsequently in different subjects. However these subjects were not requirements for our course. Thus we decided to use a more Image Oriented Systems, namely FIJI [9]. As its name implies, FIJI Is Just Image J, it has the additional advantage of having a well known and accessible API. So, FIJI can be used as prototype builder.

The Curricula core is not covered by theses topics, and the approach we take is a superficial one if it is compared with the core. However the subject can be seen as a motivator to pursue deeper math courses and later an Image Processing course as required by the Computing Curricula.

HANDWRITTEN DOCUMENTS

The Argentinean Historical Heritage Project [10] goal is planning and implementing a digital preservation strategy, converting and making accessible cultural heritage materials. The project belongs to the Instituto de Historia Argentina y Americana Dr. Emilio Ravignani which belongs to the "Facultad de Filosofia y Letras" from the University of Buenos Aires. The Institute is the guardian of several historical documents archives spanning from the 18th century to our days.

Collections were donated by notary publics, lawyers, governors or their relatives. There are ancient facsimiles and unique handwritten books. We were engaged in such project since 1996.

We decided to create "preservation-quality" digital masters-scans from the originals. These masters-scans are rich enough to use for several purposes and are created to obviate the need of rescan the original. We maintained additional information related to the digital nature of the preservation in order to extend the useful life of information resources [11] and to deal with the digital preservation until the problem of digital longevity will be resolved.

From the full quality preservation image several publication images from each original page are built. The different publication images are processed and enhanced accordingly to different goals.

One of such goals is generating an Internet access copy offering good reading quality, some of them watermarked, minimizing communication costs (to the low standard speeds of our country's links).

IMAGE PROCESSING

The problems of image processing Image processing problems are divided by standard convention into four classes:

1. Image enhancement: improving the image either for human perception (making the image "look better"), or for autonomous machine interpretation. This may include enhancing the contrast of a poorly developed image, or of enhancing the edges.
2. Image restoration: undoing the damage done to an image by a known cause. This may include removing noise, blurring, or optical effects.
3. Image segmentation: isolating aspects of the image: removing an object from the background; counting objects, etc.
4. Image compression: reducing the size of an image file.

Considering an image as an array, algorithms for processing images fall again into several classes:

1. Point operations where the output value depends only on the single corresponding input value.
2. Neighborhood operations where the output value depends on a neighborhood of values in the input image,
3. Transforms where the entire matrix is transformed and operated on as one large block.

Having access to a full quality image and a compressed version (i.e. *tiff* and *jpeg*) of the same image, students were able to understand the differences between processing a raw, information full image and a compressed version with non perceptual features deleted.

The first two classes can be explained with very little and simple math. The transforms are only explained in a surface approach and showing some examples.

As an example of our approach, we will develop how we use an ancient document (1811) to introduce a point operation such as Binary Conversion.

USING HISTOGRAMS

The following figure depicts a letter from 1811. Some prominent neighbor asks an attorney for a job for his brother.

The scanning was done to a high resolution TIFF [12] . The **Tagged Image File Format** (abbreviated **TIFF**) is a file format for storing images, including photographs and line art. As of 2009, it is under the control of Adobe Systems [14]. Originally created by the company Aldus [15]. TIFF was originally created as an attempt to get desktop scanner vendors of the mid-1980s to agree on a common scanned image file format, rather than have each company promote its own proprietary format. In the beginning, TIFF was only a binary image format (only two possible values for each pixel), since that was all that desktop scanners could handle. As scanners became more powerful, and as desktop

computer disk space became more plentiful, TIFF grew to accommodate grayscale images and later color images

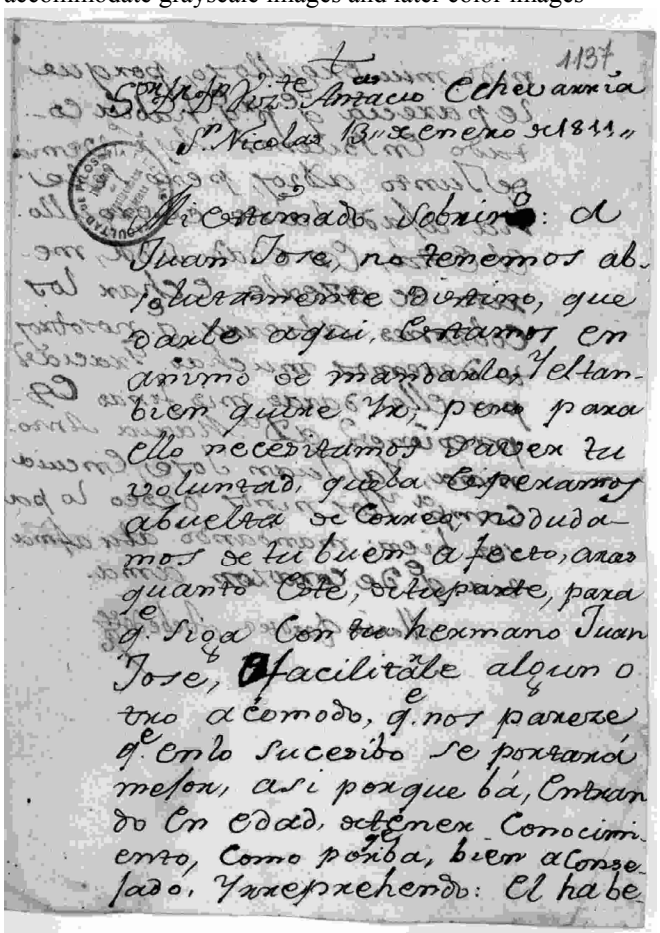

FIGURE 1
A LETTER FROM 1811 SHOWING BACK TO FRONT INTERFERENCE

In this document, a defect called “Back-to-front”, “show-through”, or “bleeding” ifs shown. These are the names given to the overlapping interference whenever a document is written (or printed) on both sides of a translucent paper. Such interference makes more difficult, if not impossible, document transcription and binarization.

Ad hoc algorithms had been developed for correcting this defects and an assessment of these methods is presented in [15].

Our aim is to use the two images (jpeg and tiff) to introduce some image processing material. Both images look the same (that's how the jpeg algorithm is developed) and because of editing limitation only one of them is reproduced in the paper.

A histogram of the jpeg image is shown in figure 2 and from the tiff image in figure 3. In the histograms students can see that some information (non perceptible) present in the raw tiff image is not present in the loosely compressed jpeg. They can also observe that the pixel count is the same for both images.

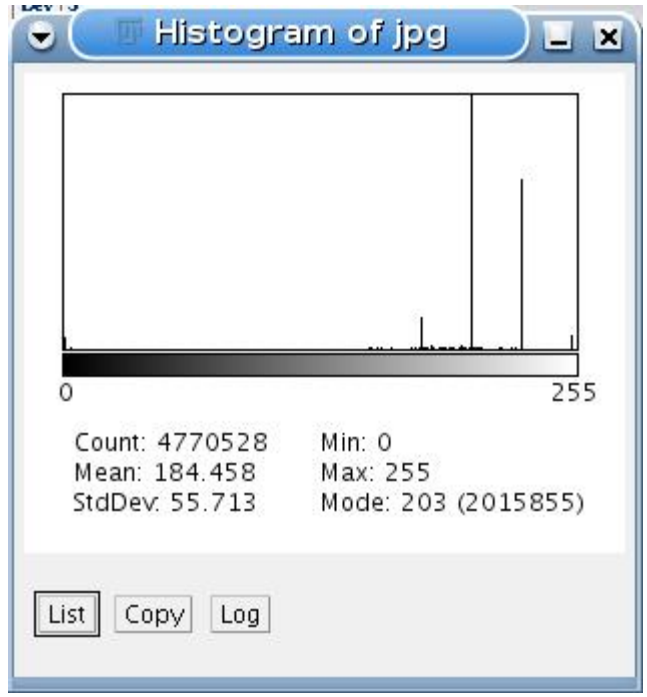

FIGURE 2
HISTOGRAM OF THE JPEG IMAGE

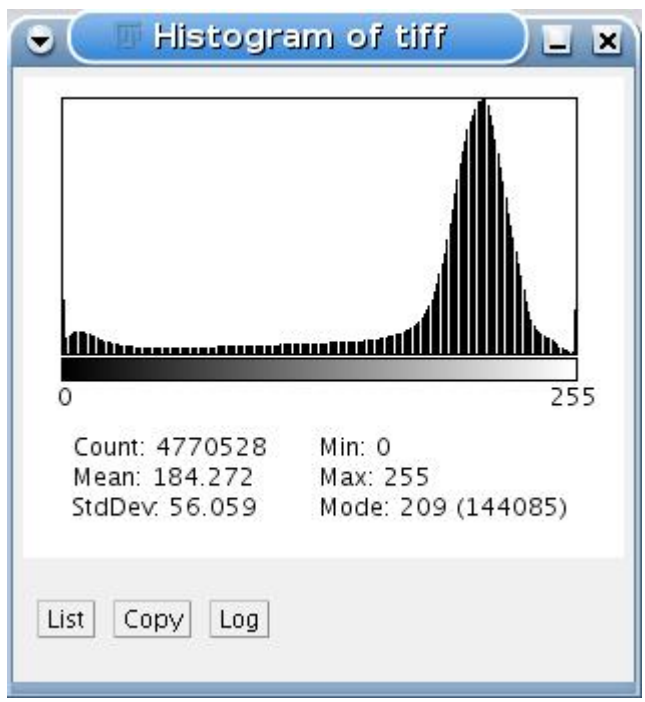

FIGURE 2
HISTOGRAM OF THE TIFF IMAGE

In order to clean the back-to-front interference, we ask our students to define a threshold and binarize the image. Working with the jpeg image they found a threshold in a pixel value of 50 (0-black, 255-white) leading Fig 3

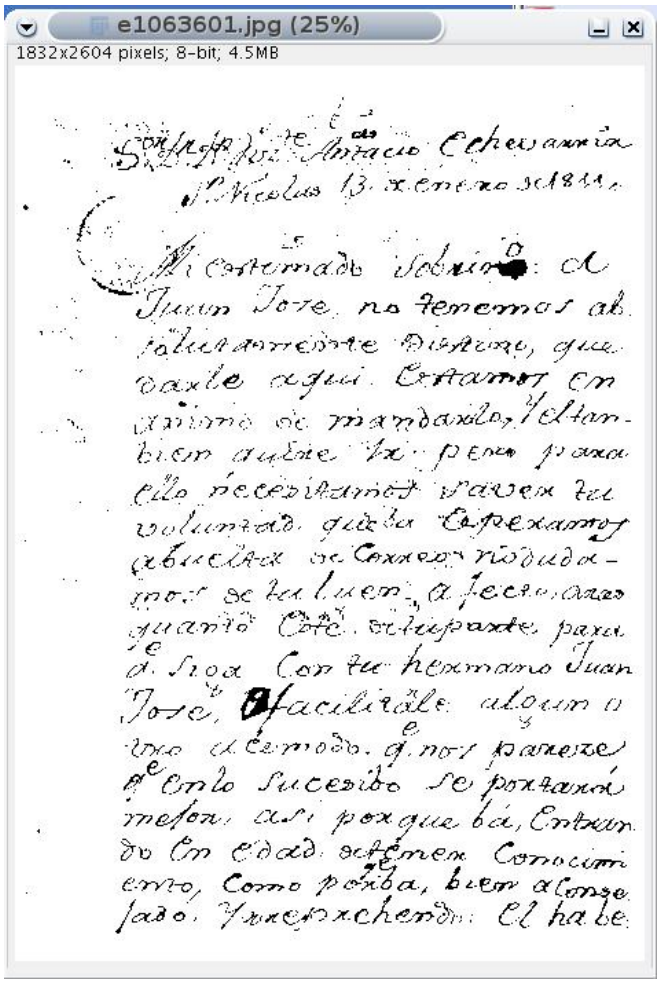

FIGURE 4
THE JPEG VERSION OF THE LETTER AFTER BINARIZING WITH A THRESHOLD OF 50

The binarizing operation in FIJI is easily done using the control of Figure 5.

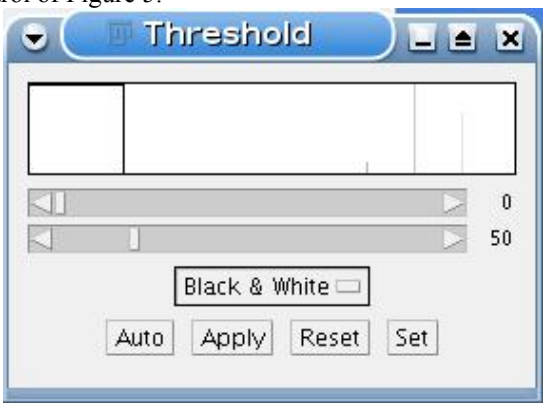

FIGURE 5
BINARIZING THE JPEG VERSION

For the tiff version, a better result can be achieved with a lower threshold

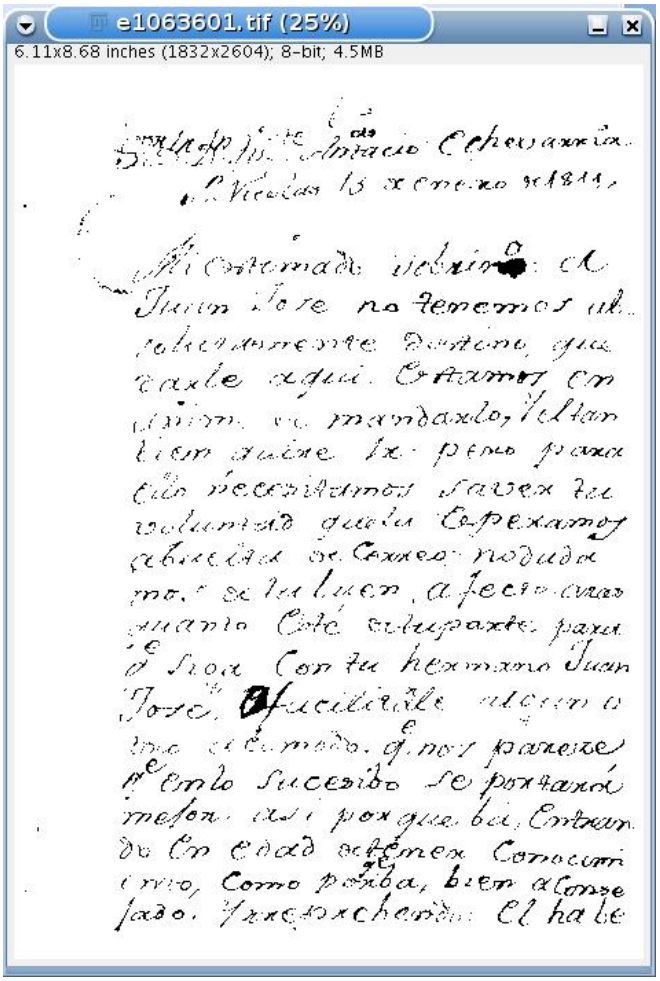

FIGURE 6
THE TIFF VERSION OF THE LETTER AFTER BINARIZING WITH A THRESHOLD OF 23

The use of information present in the tiff but not present in the jpeg can be seen in the control

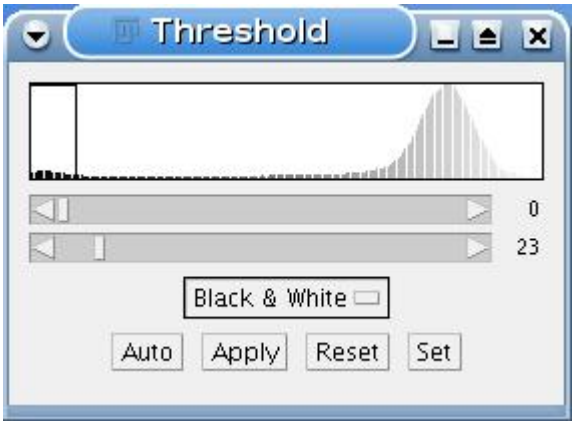

FIGURE 7
BINARIZING THE JTIFF VERSION

A comparison of both versions and of a binarized jpeg version using a pixel value of 23 is shown in figure 7

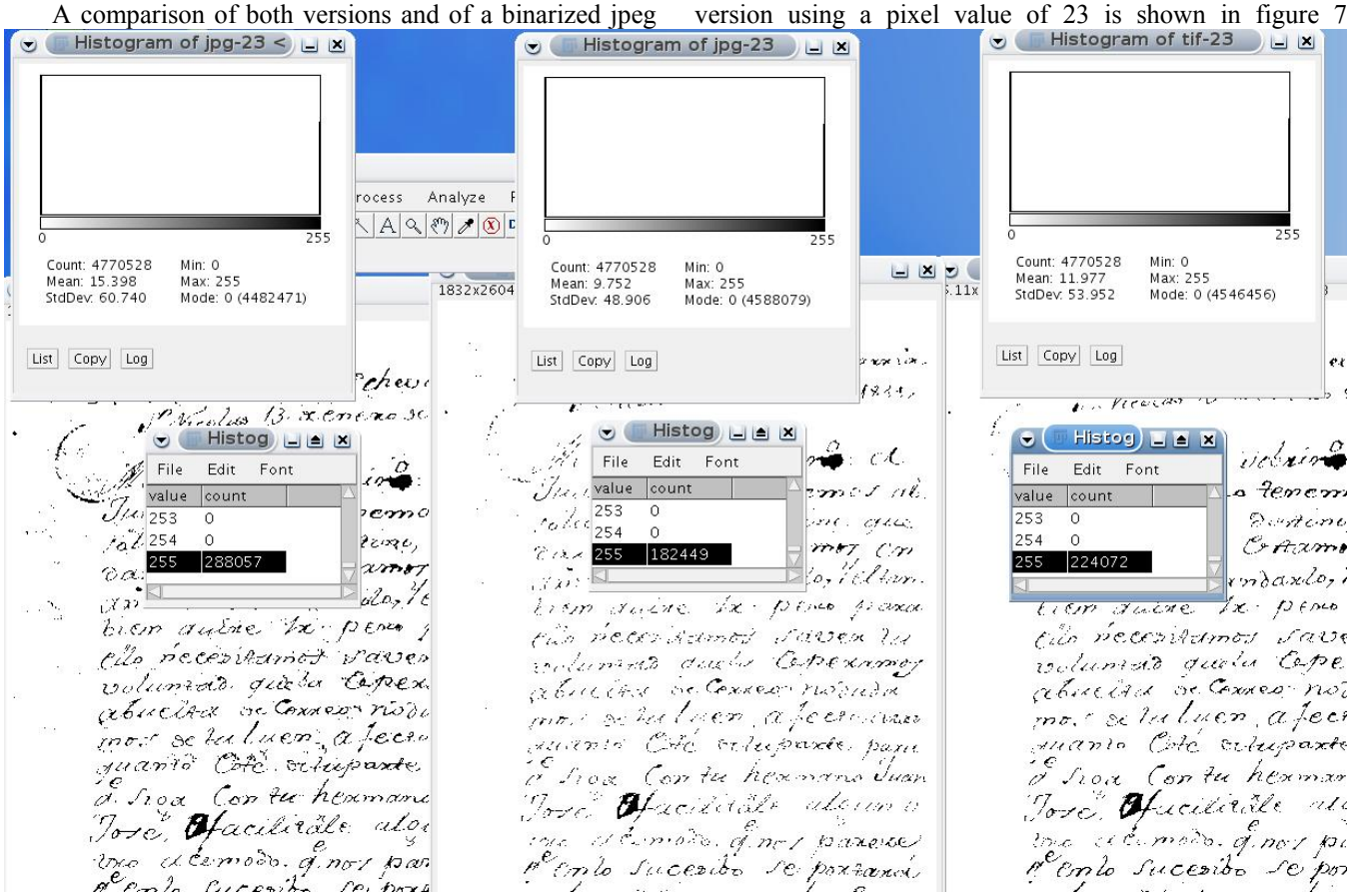

FIGURE 7

COMPARISON OF THE BINARIZED VERSION OF THE LETTER

Of course, the histograms do not present new information, but the statistics (mean and count value for black pixels) allow for some discussion with the students.

LABORATORY AND ASSESSMENT

The laboratory focuses mainly in the effect of the algorithms on different images using different parameters, as shown in the preceding example. It allows us to discuss the meaning of the different values shown in the statistics and to put some figures to the somewhat abstract concept of image quality.

Assessment is based partly on laboratory work and in a final exam. We are careful not to use more than some elemental arithmetic operations in the labs and in the exam. Students are encouraged to use different techniques and to explore the different possibilities of several image processing programs. They are allowed to use other image processing software, including general purpose photo programs such as Picasa or iPphoto. For example some students used background subtraction to address the back -to-front interference in documents.

In the final exam they have to show some project, explaining the way they processed it and relating original and resulting image to some figures, which can be found using the software.

REFERENCES

[1] A. McAndrew. Teaching image processing using minimal mathematics. In T. Greening and R. Lister, editors, *Fifth Australasian Computing Education* Conference (ACE2003), 2003.

[2] Greenberg, "Image processing for teaching: transforming a scientific research tool into an educational technology," *Proceedings of the 1996 international conference on Learning sciences*, Evanston, Illinois: International Society of the Learning Sciences, 1996, pp. 559-559.

[3] R. Wicentowski and T. Newhall, "Using image processing projects to teach CS1 topics," *Proceedings of the 36th SIGCSE technical symposium on Computer science education*, St. Louis, Missouri, USA: ACM, 2005, pp. 287-291.

[4] ACM IEEE *Computer Engineering 2004.*

[5] ACM-IEEE *Computer Science Curriculum 2008*

[6] http://www.matworks.com

[7] http://www.gnu.org/software/octave/

[8] http://www.r-project.org/

[9] http://pacific.mpi-cbg.de/wiki/index.php/Main_Page

[10] M. Feldgen, O. Clúa, F. Boro, and J.J. Santos, "Argentinean historical heritage project," *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, Portland, Oregon, USA: ACM, 2002, pp. 392-392.

[11] *Conway, P. The relevance of Preservation in a Digital World. Northeast Document Conservation Center, Technical* Leaflet, Section 5 Leafllet (Winter 1999), 1-10

[12] Tagged Image File Format - Wikipedia, the free encyclopedia.http://en.wikipedia.org/wiki/Tagged_Image_File_Format

[13] Aldus/Microsoft, "TIFF," *1) Structure*, Aldus Corporation and Microsoft Corporation, 1999.

[14] A.D. Association and Adobe Developers Association, "TIFF," *Section 7: Additional baseline TIFF Requirements*, Adobe Systems Incorporated, 1992, p.p. ;26.

[15] R.D. Lins and J.M.M.D. Silva, "A quantitative method for assessing algorithms to remove back-to-front interference in documents," *Proceedings of the 2007 ACM symposium on Applied computing*, Seoul, Korea: ACM, 2007, pp. 610-616