

Application of Hyperspectral Imaging to Forensic Document Examination Problems

R. Brent Ostrum

Canada Border Services Agency, Forensic Sciences Division, 79 Bentley Avenue, Ottawa, Ontario, Canada K2E 6T7

(Developed from a poster presented at the 63rd ASQDE Annual Meeting, Montréal, QC, 2005)

Imaging spectroscopy is a powerful tool used extensively in many different remote-sensing applications. The parallel to applications in forensic document examination is exact, but the potential of this method has not been fully explored. A customized imaging system based on electronically tunable filters was used to generate hyperspectral image data from ink samples known to be difficult to differentiate using other spectral methods. The potential of hyperspectral imaging (HSI) as a tool for forensic document problems was evaluated specifically for ink differentiations and the decipherment of obliterations on questioned documents. In this study, various HSI methods were used to show differences between test inks. The most effective results were seen in plots of spectra response, while other methods (such as pseudo-color images of select bands) proved to be ineffective. Select results were also compared to those obtained using a VSC-2000 and found to be as good or better. The basis for a fully automated HSI analysis system was developed.

Introduction

Imaging spectroscopy has been used in remote-sensing applications by the military, exploration geologists, geobotanists, as well as many other fields. Multispectral imaging (MSI) and hyperspectral imaging (HSI) are 2 common forms of imaging spectroscopy. MSI and HSI can be defined as any process that produces a multivariate image where "... [a] multivariate image is a 3-way array, a stack of congruent images, with each 2-way image in the stack measured for a different 'variable,' e.g., wavelength" (Geladi and Grahn 1996).

The distinction between MSI and HSI imaging is mainly one of resolution. MSI methods are typically based on a limited number of bands (or channels) of data; more than 3 but generally less than 10. MSI bands are usually noncontiguous so that the spectrum is divided into discrete regions that are not adjacent. HSI is different in that the data consists of dozens or hundreds of bands of spectral information for each pixel. Most researchers do not consider the method to be HSI until there are at least 30 bands of data. Furthermore, while the sampling is still discrete, the analysis generally covers a broad range of the spectrum with no gaps or discontinuities.

A multivariate image is a 3-dimensional (3D) data set that can be considered in 1 or 2 fundamental ways. The 1st is as an array of spatially resolved individual spectra whereby every pixel in the image contains an entire spectrum for the corresponding physical point in the original sample. Alternatively, the data may be viewed as a stacked series of spectrally resolved images. In either representation, the 3D data can be treated mathematically as a single entity (Figure 1).

All multispectral or hyperspectral data is simply a multivariate data array. As such, the data may be analyzed or processed like any other multivariate image. An excellent reference is Geladi and Grahn's *Multivariate Image Analysis* (1996). As with any complex statistical/data analysis, the choice of appropriate method(s) depends upon both the nature of the underlying structure of the data and the purpose of the analysis (i.e., the desired outcome of the processing or what the user is trying to achieve).

Multivariate image analyses include various preprocessing operations (blurring, median filtering, etc.), as well as methods like principle component analysis (PCA), clustering techniques, pseudo-color display of key bands, band subtractions, and spectra extractions. Some of these

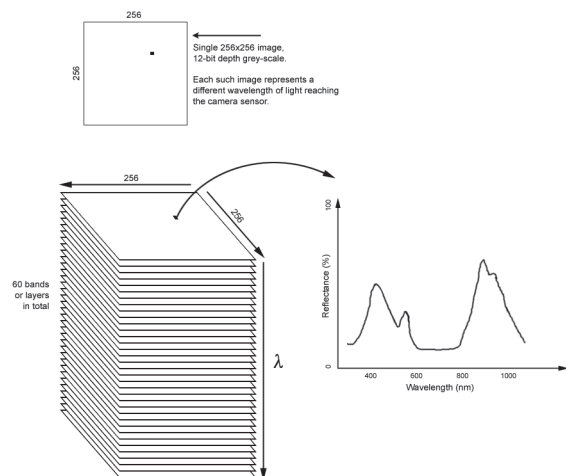


Figure 1. Individual images (top) represent the spectral response of the document at any single illumination setting. These are stacked (bottom left) to form the imagecube. Data can be analyzed as a 3D matrix or, alternatively, viewed pixel-by-pixel to see a spectrograph for any given point in the document (bottom right).

reduce the data set to a more manageable level, while others enhance trends or details hidden in the data. Some simply present the data in a more easily interpreted fashion. This study did not attempt to explore all of the options that may be applied to this type of data. Instead, focus was placed on a few methods outlined below.

Traditional imaging options, such as using the Foster + Freeman Video-Spectral Comparator (VSC) or the Projectina Docucenter, can be very effective for differentiating inks and/or deciphering obliterations. However, these methods do not always work (Ibrahim 2000, 2002). These processes can also be subjective in that the examiner must select the settings that provide the best visual results. Finally, user interaction and control are required which can be both time consuming and intensive.

HSI has been used effectively in art history and conservation for some time to study works of art, including paintings and murals. Recent work by Attas (2003, 2004) addresses problems drawn from the conservation discipline very similar to those encountered by document examiners. Those particular studies explored the use of HSI techniques to assess IR reflectance properties for both artwork and writing on older manuscripts. Among other things, the results showed that high-

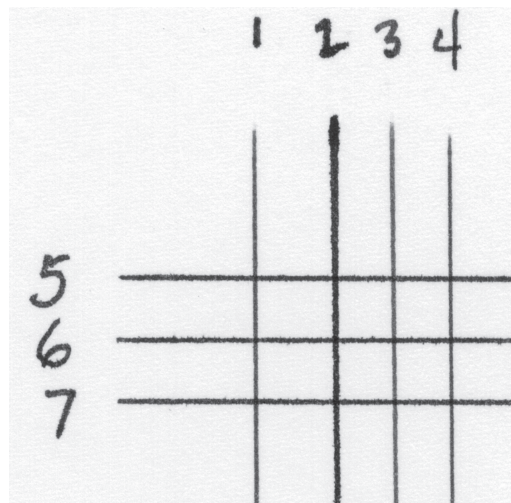


Figure 2. Test Sample #1 showing under normal lighting conditions (~ 1:1) a grid consisting of lines drawn on paper using 7 different blue ink pens.

resolution spectroscopic imaging can be useful for the enhancement and interpretation of faded ink writing on old manuscripts.

As an alternative to traditional special-lighting options, HSI has several appealing attributes, including the ability to discern and evaluate extremely subtle differences in ink spectral response, numeric analysis of such differences, potentially objective evaluation of results, and automation of the imaging procedures. In particular, the possibility of using powerful statistical methods for the analysis of potential ink differences is very appealing. Various forms of numerical analyses are discussed at length in Mansfield, et al. (1997, 1999) and Landgrebe (1998, 1999).

In general, hyperspectral images from real-world scenes tend to be very complex combinations of different components. By comparison, most documents are simple constructs having a background element and a small number of visible elements. Often, these elements will be relatively discrete and, hopefully, separable using HSI methods. In other words, HSI for questioned document purposes should be relatively easy from a processing point of view.

The overall purpose of this study was to assess HSI as a potential tool for questioned document examinations in the area of ink differentiation and



Figure 3. *Sample #2* consisted of the word “test” written on plain paper using ink 1 (Sanford Uni-ball Fine) on the left, and ink 4 (Merangue Roller refill RBRF-BKB) on the right. Ink 4 was used to obliterate ink 1 (2nd image from left) and ink 1 was used to obliterate ink 4 (image on right).

obliteration decipherments. The specific goals were to see if HSI techniques could be used to (1) enhance the visual interpretation of ink differences and/or (2) measure spectral differences in inks.

Materials and Methods

Materials consisted of:

Pens (blue inks):

1. Sanford Uni-ball Fine
2. Zebra Zeb-Roller 0.7
3. Waterman Rollerball Refill Fine
4. Merangue Roller refill RBRF-BKB
5. Staedtler Liquid Point
6. Staedtler Liquid Ball
7. Roundedge Grand & Toy 99653

Paper substrate: Weyerhaeuser First Choice Premium Ultra White (96) plain paper; 90 gm/m²

Reflectance Standard Target: Spectralon® target (>98% reflectance for 250 nm through 1800 nm)

Hardware:

1. Camera: Q-imaging PMI-1400 programmable digital camera (Kodak KAF scientific grade sensor, 1317x1035, 6.8x6.8 micron sensors) binned 2x2 to give raw images at 658x518 with 12-bits per pixel
2. Camera lens: Nikkor 50 mm 1:1.2
3. Barrier filter: Cambridge Research and Instrumentation (CRI) Varispec™ liquid-crystal tunable filter (LCTF) for Near-IR 650-1100 nm (SNIR model)
4. Light source: 2 Leica 7v Quartz-Halogen lamps
5. Foster + Freeman VSC-2000

Software:

1. Q-imaging PMI-Capture utility for image acquisition (camera control)
2. Matlab® Version 7.0.4, Release 14 (with Image Processing and Statistics version 5 Toolboxes installed), The Mathworks, Inc.

Two test samples of inks were prepared on plain paper. *Sample #1* was created using the 7 blue pen inks listed above. These inks are known to be difficult to differentiate using a Foster + Freeman VSC4c. The ink lines were drawn onto the paper in a grid pattern approximately ½" x ½" (1.5 cm²) in size (Figure 2).

Based on test results from the *Sample #1*, ink 1 (Sanford Uni-ball Fine) and ink 4 (Merangue Roller refill RBRF-BKB) were selected as being particularly difficult to differentiate; therefore, they were used to prepare *Sample #2*. The 2nd test sample consisted of the word “test” written on plain paper using pens containing ink 1 and ink 4. The “test” entries were then manually obliterated using the other pen. These entries were then analyzed to see if the obliteration could be deciphered (Figure 3).

As this procedure is based upon differences in reflectance of light at various wavelengths, it was important to minimize the presence of anomalous shading or tonal artifacts. One potential source of shading was shadows caused by strong directional lighting and/or deep writing impressions. It was decided to control for this by (1) creating all writing samples using minimal writing pressure, and (2) using proper copy lighting designed to minimize shadows. Thus, the lines in *Sample #1* were drawn using a straightedge (with care taken to use minimal writing pressure). Similarly, both the writing and obliterations in *Sample #2* were executed using light writing pressure. The obliterations consisted of 4 repetitions of strokes in different directions. The obliterations were subjectively assessed as being effective in hiding the original entry. Finally, a basic copy lighting arrangement was used to minimize shading that could result from deep indentations. This consisted of 2 semi-diffuse light sources aimed onto the surface of the paper at roughly 45° from opposite sides. It might be difficult to address this issue in a “real-world” examination. Hence, the

topic is worthy of additional research. However, this study did not explore the issue beyond an attempt to minimize the effect through the controls described above.

Basic image acquisition was achieved using a dedicated software utility for the camera and a manual control box for the LCTF filter. The camera sensor was binned 2x2 to produce 648x516, 12-bit grayscale images. Binning is a process that allows the electrical charge from 2 or more sensors to be combined on the CCD chip prior to readout. Forty-six filter settings were used. Both ink test samples and Spectralon reflectance standard were imaged at each setting. These raw data images were saved in uncompressed 16-bit TIF format, although the image data was only 12 bits.

The ink samples and the reflectance standard were illuminated using 2 strong, broadband (full-spectrum) light sources positioned for copy lighting (45° on opposite sides) aimed to produce a relatively diffuse lighting. The reflected light reaching the camera was band-pass filtered using the LCTF. At each wavelength a grayscale image was captured. A total of 46 bands of information were recorded for filter settings from 650 through 1,100 nm in 10 nm increments.

Images from the test samples and the reflectance standard must be acquired under the same conditions. To help ensure that the test sample did not move during the imaging process (thus allowing spatial superimposition of the images), a special holder was created so that the reflectance target could be moved in and out of the imaging position without affecting the position of the test sample during image capture. The individual pixel values in each captured image were 12-bit intensity readings (i.e., maximum of 4,096 grey values) representing the reflectance characteristics of the document at a given filter setting (i.e., in a given band). To be precise, the raw image data obtained in this imaging procedure represented the reflectance of incident light at each of the wavelengths passed by the LCTF.

Although the data could be analyzed with no adjustment, most HSI processing is performed on absorbance data. In keeping with this, each of the reflectance images was converted to an equivalent absorbance image using a method similar to that described by Attas (2003, 2004). The conversion formula used was $-\log(I_s/I_R)$; whereby I_s is the intensity of a reflectance image pixel and I_R is the intensity of the corresponding pixel from the reference standard image captured at the same wavelength.

Data Pre-Processing

Primary processing of the data was done in the Matlab numeric computing environment on a computer running Windows 2000. Matlab provides extensive capabilities for image processing and numerical data analysis; it is particularly well-suited for arrays and matrices of a multidimensional nature.

Each set of images was preprocessed as follows:

- Individual images were loaded into Matlab as 2-dimensional data array with 12-bit entries. The images were stacked to create a 3-dimensional array (i.e., the imagecube as described above and depicted in Figure 1).
- The bands in the test (ink) imagecube were balanced with the equivalent reflectance standard data (band by band) and converted to absorbance values using the $-\log(I_s/I_R)$ formula.
- The resulting data set was cropped to a 256x256 area of interest (AOI) to remove lens vignette.
- The absorbance imagecube data was then saved in a multiband format.

Note that standard imaging procedures (integer based) are not acceptable for this type of processing because rounding errors will result. To minimize potential rounding errors, data was converted to double precision floating point and processing performed in this mode.

Analyses

Matlab provides 2 useful functions for multichannel/multivariate images. Specifically these are related to the importing or exporting of the data in valid Multiband format (which is the standard format for any multispectral or hyperspectral data) by way of the Multiband read and Multiband write functions (Matlab User's Guide).

There are no other built-in functions for analyzing multiband images, but the program is capable of performing any desired mathematical analysis on multidimensional data arrays so long as the user is able to create the appropriate sequence of functions. The Matlab environment also allows a user to develop such methods quickly and easily. All of the data processing in this study was done in Matlab with a macro or "m-file" developed for each purpose. Customized Matlab routines used in this study were developed using instructions provided in *Multivariate Image Analysis* (Geladi 1996).

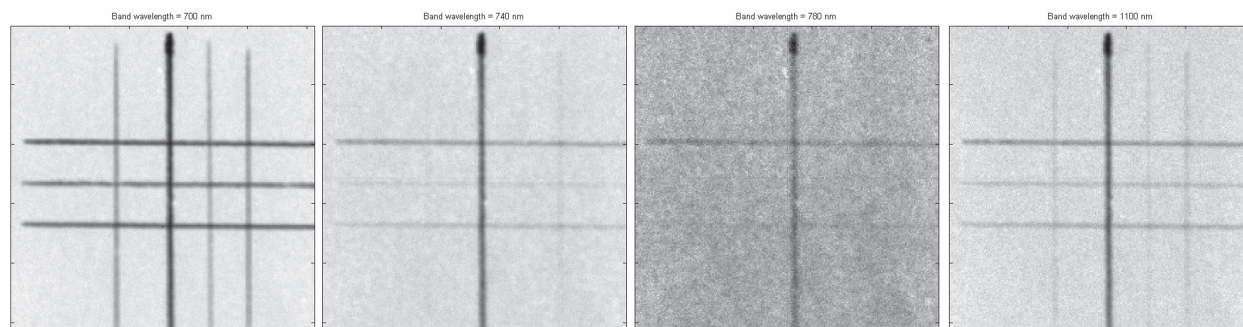


Figure 4. Examples of band images clearly showing ink differentiation beginning left: Band 6 (700 nm) Band 10 (740 nm), Band 14 (780 nm), and Band 46 (1,100 nm).

Ink Differentiation

The imagecube was processed using several standard HSI processes. Various macros were created to perform the analyses:

- Direct visual examination of individual bands—no processing (simply band-pass filtered).
- RGB pseudo-color representations of key bands.
- Extraction and plotting of pixel spectra (by user-selected locations using visible image references).
- Calculation of PCA to generate loadings and calculation of T_x score images that were examined directly. Each T-score image is a graphical representation of the data “concentrated” in factors produced using the PCA process. T-score images concentrate the meaningful data in the 1st several images (low numbers) so that most of the other images can be discarded.
- RGB pseudo-color representations of key score images.

Decipherment of Obliterated Entries

The imagecube was processed, and the PCA loadings were used to generate T_x score images. The images were reviewed, and 3 were used to create a suitable pseudo-color RGB image(s).

Results

Each type of analysis was reviewed separately, with the various results depicted visually as images whenever possible. Please note that only a few representative images are provided here as examples.

Ink Differentiation

- Images (Figure 4) depict specific bands with some visible differences in the inks. As expected, certain bands will differentiate some of the inks but not all of them.
- RGB pseudo-color representations of key bands were created, but the results were not particularly beneficial.
- One potentially useful approach for differentiating regions in an image is to take samples from each region/ink line and generate an “average spectra” for the ink. The average spectra are presumed to be representative of the region on the assumption that the sampling is sufficient and that the area being sampled is reasonably homogenous.

A Matlab routine was created in which the user manually selected points in the image for up to 8 different features (e.g., ink or paper). The user sampled up to 5 points for each feature of interest. The routine then produced an average spectra based on the points selected and plotted the results on a common axis (Figure 5). In this instance, all 7 of the ink lines were sampled 5 times each to generate a spectral plot for each ink. The paper was also sampled. The results were promising, as each of the regions/features showed visually divergent spectra. The differences are most apparent in the earlier bands (1 through 15) and again in the region from bands 17 through 45. It should be noted that this was a preliminary test of limited value. Additional testing must be done to ensure both reliability and accuracy of these results.

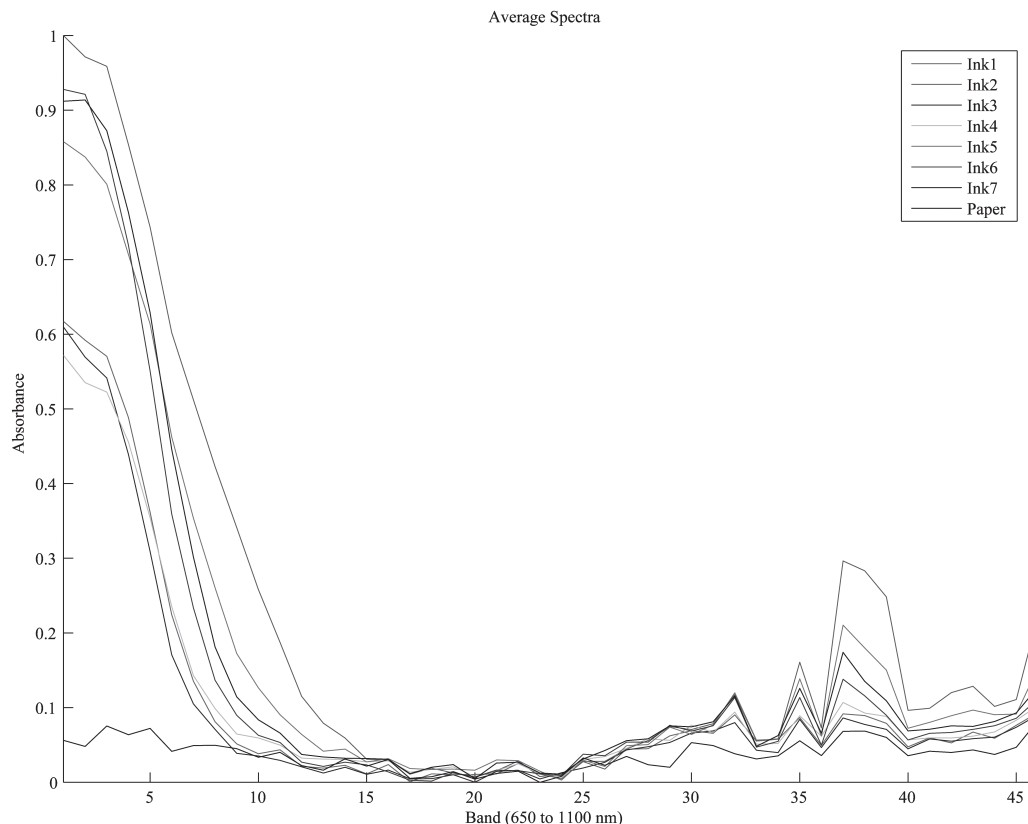


Figure 5. An average spectra based on the points selected and plotted.

A general limitation of this technique is that it is entirely dependent upon the user being able to select representative points from the image. In this instance, this was not a difficult task because of the obvious geometric layout of the samples and the resulting knowledge of where sampling could be done to best effect. In many images, it would not be as easy to select representative points for each ink feature.

A PCA was conducted. The resulting loading factors were used to create 46 “score” images, designated T_1 through T_{46} . As noted earlier, T-score images concentrate the meaningful data in the 1st several images (i.e., low numbers), with the result being that the data is concentrated into the 1st few T-score images (Figure 6). These images demonstrate differences between inks. Any given image does not likely differentiate all inks, but in this instance, all of the inks can be visually differentiated by reference to the 1st 6 T-images, T_1 through T_6 .

- d. Another common visualization method is to use up to 3 of the T-score images as channels in a RGB image. The result is a pseudo-color image that may allow differentiation of several features from one image (Figure 7). In the example images shown in Figure 6, score images T_1 , T_2 , T_3 , T_5 , and T_6 were combined to produce 3 pseudo-color images.

Decipherment of Obliterated Entries

Various analyses were used with the obliteration test samples. The most effective HSI result was obtained using T-score images T_2 and T_3 together as an overlay (Figure 8). Results from the VSC-2000 are provided here to show an alternative method available in many laboratories.

Discussion

Ink Differentiation

Direct visual examination of individual imagecube bands with no additional processing is roughly equivalent to simple band-pass filtering using, in this case—a narrow (10 nm) band

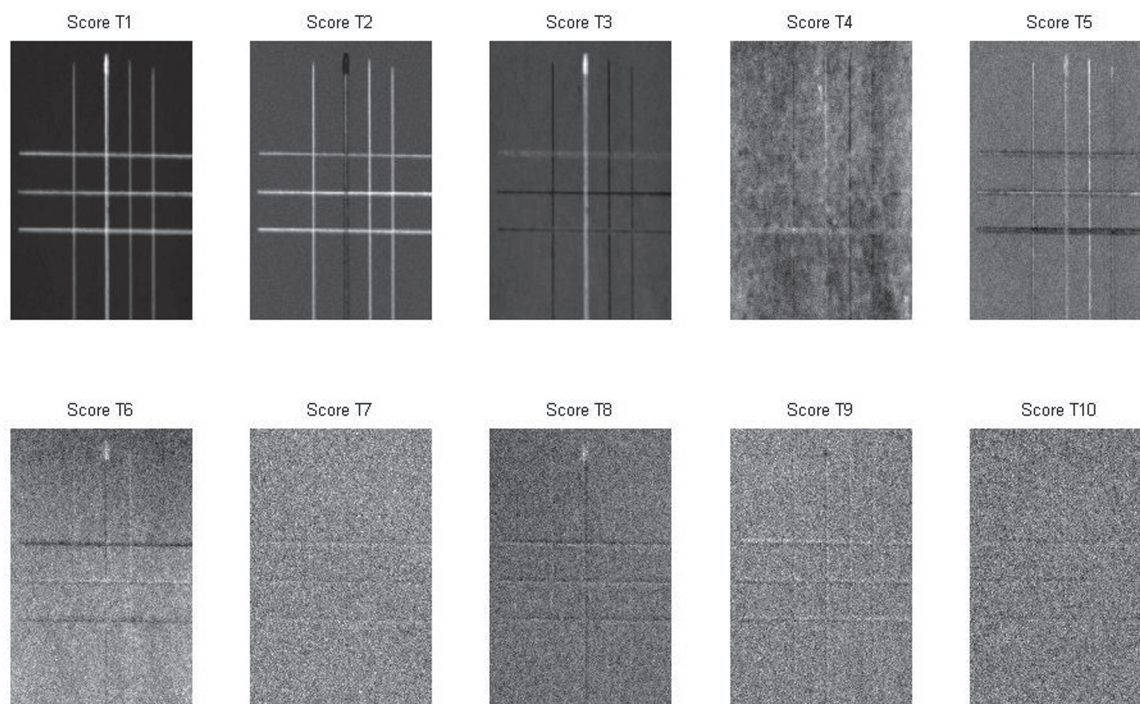


Figure 6. Data concentrated into the 1st 10 score images, T_1 through T_{10} .

pass window. This was reasonably effective in differentiating a few of the inks that can be seen in the images of bands 6 (700 nm), 10 (740 nm), 14 (780 nm) and 46 (1100 nm). By referring to several images, it was possible to distinguish some of the inks; however, it was difficult to integrate the results for the entire set. A potential solution to this is to create RGB pseudo-color combinations of 3 different bands or T-score images. Pseudo-coloration is a very useful imaging technique that can enhance subtle tonal differences. However, in this instance, the pseudo-color images based upon band images did not help in distinguishing the inks.

When specific regions of interest can be selected on the original document (i.e., the ink lines are well defined) it may be possible to select representative locations and generate spectra representing the material present at those locations. In the above examples, the spectra obtained had subtle but significant differences making this an effective method for differentiating all of the inks.

PCA is a powerful statistical method for the reduction of a large data set to a more manageable size and format. While it is possible to assess loading factors directly (i.e., by viewing a table of the numeric results), they are often used to create

score images (in the present example, T_x where $x=1$ to 46). The visual results were interesting and did provide better discrimination than the original band images. The useful information was concentrated into the 1st 6 score images. The pseudo-color images derived from these T-score images were interesting and provided slightly better differentiation than some of the other visualization methods.

As a side note, score images can be very useful for data visualization, but additional processing should not be done directly to these images as a general rule. The data comprising these images is greatly reduced from the original data set. To display the data set as an on-screen image (or as a printed version), the data must first be scaled from 12-bit (0 to 4095) to 8-bit (0 to 255) and rounded to integer values. If additional processing is required, the original imagecube data should be used and new images generated as appropriate.

Decipherment of Obliterated Entries

The inks used in this test (pens 1 and 4) were selected as being difficult to differentiate based on the ink differentiation findings. As a result, this was a relatively difficult challenge. Decipherment was good for the test with ink 4 obliterating

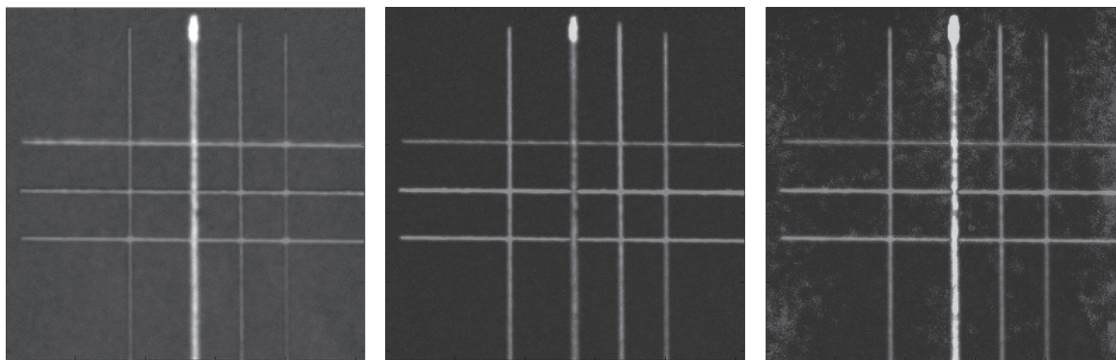


Figure 7. Pseudo-color representations using T1, T2, T3, and T5: left represents T1, T3, T5; center T2, T3, T5, and right T2, T3, T4.

ink 1. The test with ink 1 obliterating ink 4 was more difficult, and the entries could only be partially deciphered. Note that images viewed on screen are a much better representation than the printed versions provided here.

For comparison purposes, the obliterations were also examined using a VSC-2000. It was expected that both entries would be decipherable, and this proved to be the case (VSC settings: 400-540 nm incident light, 645 nm longpass). For these particular inks, the VSC images were clearer and easier to decipher than the HSI images, especially for the test with ink 1 obliterating ink 4. However, in order to achieve this decipherment, it was necessary to override the automatic VSC functions. This is not a difficult thing to do, but it requires appropriate knowledge of the operation of the device and some patience. The VSC imaging process involved extensive experimentation and took about 45 minutes total time.

In contrast, the HSI images took a total of 20 minutes to prepare (about 10 minutes for acquisition and another 10 to process) and required almost no user interaction. With additional development, the acquisition time could be reduced considerably. However, while the visible end results were promising, they were definitely not exceptional. Before this particular method can be considered an effective option, additional development and improvement will be required.

Conclusions

Overall, the test results demonstrate that standard HSI/data processing options may be another effective tool for forensic document examiners when faced with problems involving the

differentiation of inks or the decipherment of obliterations. It was possible to demonstrate differences in each of the inks in the 1st test sample; particularly with the spectra plots. Although the results were not quantified or assessed numerically, this is a possible benefit of the method since such techniques are often applied in other similar HSI applications.

The methods did decipher the obliterated entries in the 2nd test sample, but not particularly well. The most significant benefit at this time is that the HSI procedure did not require much user interaction or control, and it was easy and quick. It must be emphasized that the present study was very limited, and the results should not be extended too far. The basic techniques and methods for HSI are well established in various fields, but additional work must be done looking specifically at questioned document related materials and scenarios. For example, additional studies should examine the effect of writing pressure (i.e., depth of writing indentation and possible resulting shadows) together with different lighting options that might be used to overcome any negative effects.

The equipment being used in the author's laboratory requires manual control and adjustment of the imaging equipment during image/data acquisition. It is anticipated that some degree of user interaction will always be required in the processing phase. However, it should be possible to fully automate the acquisition process as well as most of the more basic analyses. This will enhance the objectivity of this type of analysis and reduce the time commitment for the analyst/examiner.

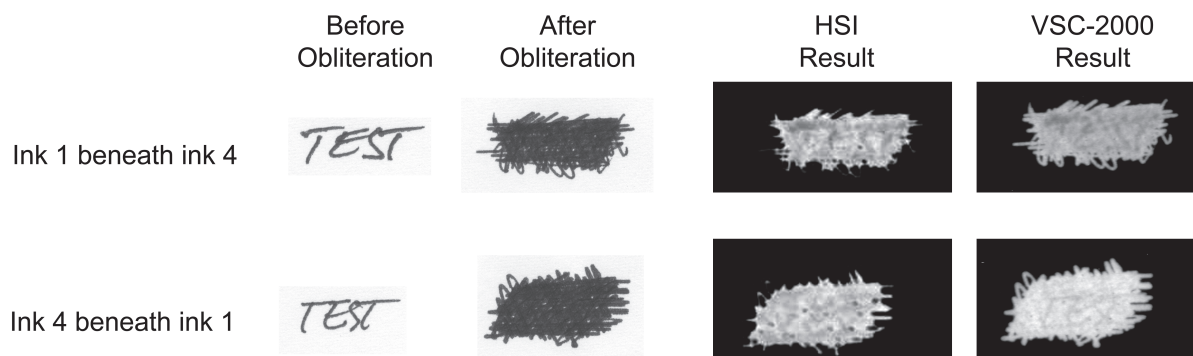


Figure 8. Comparison chart with obliteration test samples showing HSI results using T-score images T2 and T3 together as an overlay and VSC-2000 results.

References

- Attas, E.M. (2004). "Enhancement of Document Legibility Using Spectroscopic Imaging." *Archivaria*, Vol. 57, Spring, pp. 129-144.
- Attas, E.M.; Cloutis, E.; Collins, C.; Goltz, D.; Majzels, C.; Mansfield, J.R.; and Mantsch, H.H. (2003). "NIR Spectral Imaging in Art Conservation: Investigation of drawing constituents." *Journal of Cultural Heritage*, Vol. 4, pp. 127-36.
- Attas, E.M.; Cloutis, E.; Collins, C.; Goltz, D.; Majzels, C.; Mansfield, J.R.; and Mantsch, H.H. (2003). "Spectral Imaging in Art Conservation: A new tool for materials investigations." *Leonardo*, Vol. 36, pp. 304-307.
- Gat, N. (2000). "Imaging Spectroscopy Using Tunable Filters: A Review." *Proc SPIE, Int Soc Opt Eng.*, Vol. 4056, pp. 50-64.
- Geladi, P. and Grahn, H. (1996). *Multivariate Image Analysis*. John Wiley and Sons: West Sussex, England.
- Hahn, B. (2002). *Essential Matlab for Scientists and Engineers*. Elsevier: Burlington, MA.
- Ibrahim, S. (2000). "Ink Differentiation Using Liquid Crystal Tunable Filters and a Forensic Light Source." Presented at ASQDE Conference, Ottawa, ON, Canada.
- Ibrahim, S. (2002). "Spectral vs. Chemical Ink Differentiation." Presented at ASQDE Conference, San Diego, CA.
- Landgrebe, D. (1998). "Multispectral Data Analysis: A Signal Theory Perspective." www.cobweb.ecn.purdue.edu/~landgreb/publications.html.
- Landgrebe, D. (1999). "On Information Extraction Principles for Hyperspectral Data—A White Paper." Geocomputation 99 Conference, Fredericksburg, VA, USA.
- Mansfield, J.R.; Sowa, M.G.; Scarth, G.B.; Somorjai, R.L.; and Mantsch, H.H. (1997). "Analysis of spectroscopic imaging data by fuzzy C-means clustering." *Anal. Chem.* Vol. 69, pp. 3370-3374.
- Mansfield, J.R.; Majzels, C.; Collins, C.; Cloutis, E.; Mantsch, H.H. (1999). "Near Infrared Spectroscopic Reflectance Imaging: Supervised versus Unsupervised Analysis Using an Art Conservation Application." *Vibr. Spec.*, Vol. 19, pp. 33-45.
- Matlab Image Processing Toolbox User's Guide, Version 5 (2006).