DOI: 10.1039/C4AN00961D

Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2014, **139**, 5176

Near infrared hyperspectral imaging for forensic analysis of document forgery

Carolina S. Silva,[a] Maria Fernanda Pimentel,*[b] Ricardo S. Honorato,[c] Celio Pasquini,[d] José M. Prats-Montalbán[e] and Alberto Ferrer[e]

Hyperspectral images in the near infrared range (HSI-NIR) were evaluated as a nondestructive method to detect fraud in documents. Three different types of typical forgeries were simulated by (a) obliterating text, (b) adding text and (c) approaching the crossing lines problem. The simulated samples were imaged in the range of 928–2524 nm with spectral and spatial resolutions of 6.3 nm and 10 μm, respectively. After data pre-processing, different chemometric techniques were evaluated for each type of forgery. Principal component analysis (PCA) was performed to elucidate the first two types of adulteration, (a) and (b). Moreover, Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) was used in an attempt to improve the results of the type (a) obliteration and type (b) adding text problems. Finally, MCR-ALS and Partial Least Squares–Discriminant Analysis (PLS-DA), employed as a variable selection tool, were used to study the type (c) forgeries, *i.e.* crossing lines problem. Type (a) forgeries (obliterating text) were successfully identified in 43% of the samples using both the chemometric methods (PCA and MCR-ALS). Type (b) forgeries (adding text) were successfully identified in 82% of the samples using both the methods (PCA and MCR-ALS). Finally, type (c) forgeries (crossing lines) were successfully identified in 85% of the samples. The results demonstrate the potential of HSI-NIR associated with chemometric tools to support document forgery identification.

Received 28th May 2014
Accepted 22nd July 2014

DOI: 10.1039/c4an00961d

www.rsc.org/analyst

1. Introduction

A major issue in police departments is document forgery. There are a variety of problems associated with this subject. Often, frauds occur by the misuse of blank documents by persons who are trusted by the signer or by the adulteration of an official document. Most of the reports involving document analysis to detect frauds employ destructive methodologies.[1–6] The use of nondestructive analytical methods, in cases where the authenticity of documents is questioned, is essential in litigation processes, where the evidence must be preserved.[7] For example, Lalli *et al.*[8] presented a nondestructive methodology to identify counterfeit documents at specific regions. The authors used an easy ambient sonic-spray ionization mass spectrometry technique for dating inks in two types of forgeries: (i) line crossing and (ii) superimposition. Using blue and red pens, purchased in commercial establishments, the authors acquired chemical fingerprints from the inks used to produce registers, and compared the profiles acquired from the fresh ink, accelerated aged ink and naturally aged ink on legal documents provided by the Brazilian Federal Police. Fresh and aged inks showed different mass spectral fingerprints. In addition to that, for intersecting point issue, it was possible to identify successfully which pen was used to draw the first line and which was used for the second one.

Police departments are currently using video spectral comparators. This equipment contains digital imaging systems equipped with a camera, zoom lens, and a range of viewing filters with multiple illumination sources (from UV, visible to NIR). An integral micro spectrometer measures the reflectance, transmission and fluorescence features of the sample. This equipment offers an easy way to identify altered and counterfeit documents; however, the detections are limited to univariate analysis by subjective visual techniques. As highlighted by Reed *et al.*[9] the interpretation of the data provided by this type of equipment could be improved by the use of multivariate chemometric methods of analysis.

Hyperspectral image analysis (HSI analysis) has been used for forensic purposes with different approaches. Tahtouh *et al.*[10] used HSI-middle infrared spectroscopy (HSI-MIR) to improve the localization and acquisition of fingerprints from different surfaces. Chen *et al.*[11] demonstrated the use of HSI-MIR in the

[a] *Departamento de Química Fundamental, Universidade Federal de Pernambuco, Recife, PE, Brazil*

[b] *Departamento de Engenharia Química, Universidade Federal de Pernambuco, Rua Prof. Arthur de Sá S/N, Cidade Universitária, 50740–521, Recife-PE, Brazil. E-mail: mfernanda.pimentel@gmail.com; mfp@ufpe.br; Tel: +55 81 21268729. Fax: +55 81 21267235*

[c] *Superintendência Regional em Pernambuco, Polícia Federal, Brazil*

[d] *Instituto de Química, Universidade Estadual de Campinas, Campinas, Brazil*

[e] *Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, UniversitatPolitècnica de València, Spain*

identification of latent fingermarks using score images from Principal Component Analysis (PCA) models. The authors also used a hexahydro-1,3,5-trinitro-1,3,5-triazine solution to contaminate the fingerprints and to identify the traces of this compound at fingermarks. Edelman *et al.*[12] used HSI-Near Infrared (NIR) and visible (VIS) spectroscopy associated with chemometric tools to identify the absolute and relative age of blood stains at crime scenes. Another paper dealing with HSI-NIR and bloodstains was published by Schuler *et al.*[13] who evaluated the potential of HSI-NIR to identify blood stains on different black fabrics.

Few papers discuss the use of images associated with chemometric methods as a nondestructive methodology to identify document forgery. Usually, the studies performing this kind of investigated use RGB images. Chaikovsky *et al.*[14] proposed an RGB image analysis method based in color separation using common image treatment software to separate stamp inks from pen inks. Although it is possible to separate different inks using this methodology, not every combination can be successfully identified. Kaur *et al.*[15] used DocuCenter Expert (PIA-6000) software to identify the sequence of intersecting lines in a document through its RGB image. However, this work was not successful, indicating that there is still a need for efficient methodologies to solve problems involving document forgery.

Berger[16] used UV-vis images associated with support vector machines (SVM) and Multivariate Image Analysis (MIA) techniques to discriminate between blue ballpoint pen inks in samples that simulate document forgeries of crossing lines. Several graphic instruments were evaluated by Bjoko *et al.*[17] with the aim of determining the sequence of intersecting lines. In this work, the authors produced samples combining different kinds of instruments (different types of pens, toner, ink from inkjet printings, *etc.*) to make intersecting lines and analyzed each line with a FT-MIR spectrometer coupled with an infrared microscope (range from 900 to 4000 cm^{-1}). The authors claim that they were able to determine the sequence of the intersecting lines only for the combinations with ballpoint pens and laser printers by evaluating the images at two different wavelengths.

Payne *et al.*[18] published a paper showing preliminary results of ink analysis using UV-vis, vis-NIR and fluorescence chemical images to analyze different brands of pens. Specifically, in the vis-NIR range, the authors analyzed only blue ballpoint pen inks using the PCA of hyperspectral images to discriminate among different brands of 9 blue ballpoint pens purchased. Because 4 out of the 9 pens could not be differentiated by the proposed technique, further investigation is suggested.

A recent article published in Chemical & Engineering News[19] emphasizes the importance of more precise and less subjective methods of analysis in forensic science. Several cases of forged reports and misleading results support the idea that new scientific based techniques are required to produce more reliable results.

The aim of this work was to use HSI-NIR and multivariate analysis to identify three different kinds of common forgery in suspicious documents. The analyses were conducted to evaluate the addition or modification of numeric characters on sheets of bank checks, identify obliterated texts, and the sequence of use, at the intersecting point of lines, produced by graphic instruments employed to prepare the document in question.

1.1. Hyperspectral images and chemometric tools

NIR images consist of a three-dimensional array, known as a hyperspectral cube (Fig. 1), which has spatial information (*x* and *y* axis) and a spectral dimension (*z* axis). It allows the visualization of the distribution of compounds in a sample. When unfolded to a two-dimensional matrix, these images can be treated with appropriate chemometric techniques.[20]

One of the most well-known chemometric techniques is Principal Component Analysis.[21] PCA uses latent variables to describe the original dataset variability, reducing dimensionality according to the variance related to each new uncorrelated latent variable called principal component (PC) joining T matrix, and providing information regarding how the original data interact to form each PC,[20] gathered by the loadings matrix P. In this manner, any X matrix (as such in the right of Fig. 1) can be expressed as:

$$\mathrm{X} = \mathrm{TP}^{\mathrm{T}} + \mathrm{R} \tag{1}$$

where X is the original data matrix, T gathers the PCs, P the loadings and R the residuals.

The application of PCA to images was initially named as Multivariate Image Analysis.[22,23] MIA provides score images that describe the variability of the dataset by the scores of each pixel at each PC. These new images can contain the relevant information to describe the dataset. When dealing with chemical hyperspectral images, the data dimensionality will be significantly lower because only a few components are sufficient to account for structure information related to the chemical composition of the sample.[23,24]

Usually, in hyperspectral imaging, the goal is to identify the chemical information present in each pixel. Therefore, alternative chemometric that do not impose orthogonality in the new compressed variables may help in providing more interpretable images. Multivariate Curve Resolution-Alternating Least Squares, MCR-ALS,[25] is an iterative resolution algorithm that uses initial estimate spectra as an input to find optimized pure spectra for each compound. In fact, MCR can be used for

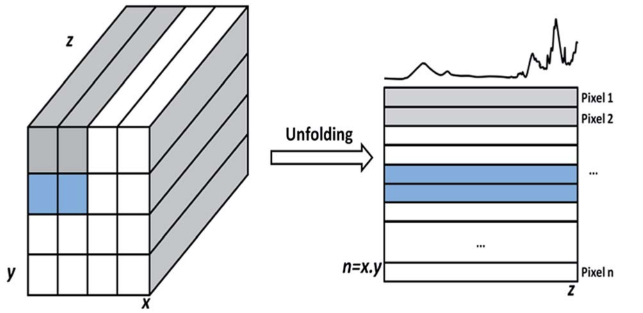

Fig. 1 NIR image as a three-dimensional array (hyperspectral cube) and the unfolding process to give to a two-dimensional matrix.

any data set that obeys bilinear models, independently of the nature of the spectral measurement. The technique works by searching for the Beer–Lambert law solutions (eqn (2)) in an iterative manner using an alternating least squares algorithm. The Beer–Lambert law states that:

$$X = CS^{T} + E \tag{2}$$

where X is the measured data matrix, C is the concentration matrix, S is the matrix that contains the pure spectrum of each compound and E is the residual matrix.[23]

However, because there are several variations of spectral behavior in image data, it is necessary to impose some constraints on the solutions to make the equation solution acceptable,[26] as well as to use some initial guesswork for the spectra. If there is no previous knowledge regarding the number of constituents in a sample (or an image), one common way for approaching this is to perform a PCA model and take a look at the number of relevant PCs.

Once the number of likely dynamics behaviors present in the image is determined, the purest dynamics in the raw data are sought using, *e.g.* SIMPLISMA.[27] In this work, we used the algorithm implemented in the software available at the multivariate curve resolution homepage.[28] When the initial guess of the pure spectra in the sample are obtained, it is possible to use them as an input to the MCR-ALS process.

2. Experimental

2.1. Samples

Ten different black pens were purchased commercially and used to produce samples. Four different types of pen were employed: 6 different brands/models of ballpoint pens (Ballpoint pen Paper Mate Kilometrica 100-BP1; Ballpoint pen Pilot BPS-Grip-BP2; Ballpoint pen Pentel RSVP BK90 Fine-BP3; Ballpoint pen Pentel WOW! BK437-A-BP4; Ballpoint pen Compactor-BP5; Ballpoint pen Bic AK13-BP6); 2 brands of gel pens (Gel Pentel EnerGel Metal Tip-G1; Gel Bic Velocity Gel-G2); 1 brand of liquid roller ball (Roller Ball Bic Grip Roller USA-RB1); and 1 brand of hydrographic pen (Hydrografic Paper Mate-H1). The pens were the most popular brand of each type. For the case of forgery by adding text, 6 additional black pens were used, as described in Section 2.1.2. All the samples were produced using A4 sulfite paper from the same batch (CHAMEQUINHO®, 75 g m^{-2} batch MGJ0L 08 12 02), except for the samples representing the falsification by adding text, which were prepared using sheets of bank checks. The samples were produced differently depending on the approach to the particular problem that was the most representative of the problems faced by the police officers in the field.

2.1.1. Forgery by obliteration. Ten different pens were used to produce a set of samples forged by obliteration. Each pen was used to write a short text on 9 pieces of paper and the remaining pens were used to scrawl over the text one week later to minimize the mixing of inks. A total of 90 samples were prepared. Fig. 2a–c show three of these samples.

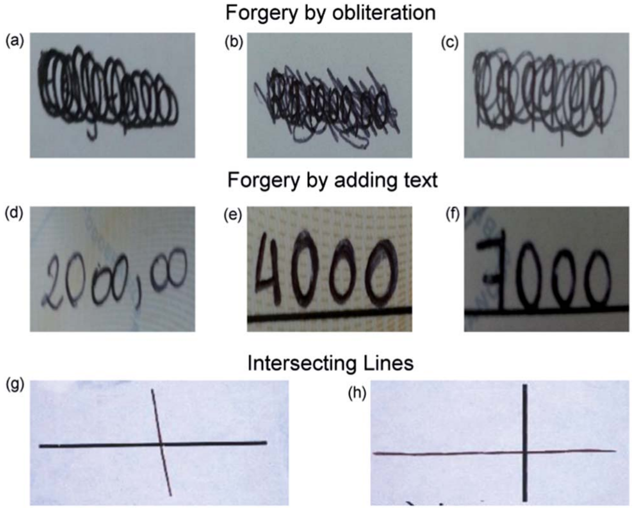

Fig. 2 Examples of prepared samples. (a), (b) and (c) simulating the forgery by obliteration; (d), (e) and (f) by adding text; (g) and (h) intersecting lines samples.

2.1.2. Forgery by adding text. Initially, the pens with the same shades of black and tip thicknesses were identified among the 10 pens employed in this study. Because the registers made with the pens required to look similar on visual inspection, the number of samples was reduced to five. To increase the data set, six more pens were purchased, similar to the ones chosen (in total 16 pens): 3 ballpoint pens (Ballpoint pen Cello Work-BP7; Ballpoint Ponta Fina Faber Castell-BP8; Ballpoint pen Uni Power Tank-BP9), 1 roller ball (Bic roller ball Triumph-RB2), 1 gel (Schneider Topball 857-G3) and 1 hydrographic (Hydrografic Staedtler Triplus Fineliner-H2). Each sample was prepared by writing a number on a sheet of bank check using one pen (pen 1). This number was subsequently modified using another pen (pen 2), which was similar to the first one. According to the criteria adopted (registers should be similar for visual inspection), twenty-two samples were used to evaluate this type of forgery. Fig. 2d–f show three of these samples (S1, S23 and S20, respectively).

2.1.3. Sequence of intersecting lines. Each one of the 10 pens was used to produce two samples, both consisting of two crossing lines: one line was drawn with a pen and the other was printed by a laser printer employing standard toners (HP Color LaserJet Black Print-toner 1). In one of these two samples, the pen line lies above the toner line, while in the other one the pen line lies below the toner line (Fig. 2g and h). The process was then repeated using a different printer and toner (HP LaserJet P2015 PCL6 Class Driver-toner 2). In this manner, 40 samples were produced, 20 pairs using 10 pens with two types of toner. To minimize mixing between toner and ink at the intersection point, the second line was added one week after the first one in all the samples.

2.2. Spectra acquisition

Near infrared hyperspectral images from each sample were acquired using the chemical imaging system SisuCHEMA from

Specim. The spectral range employed was 930–2520 nm, the spectral sampling was 6.3 nm at a spectral resolution of 10 nm. Images were acquired with a macro lens with a pixel size of 30 × 30 μm with the total number of pixels varying according to the image size. The integrity of the samples was completely preserved during all spectra acquisition.

2.3. Data treatment

Different kinds of chemometric treatments were evaluated, depending on the forgery type. The spectra of the images of all the samples were pre-processed to minimize irrelevant information and then, subjected to different analysis techniques. All chemometric treatments were performed using Matlab® R2010a. MCR-ALS was carried out by the interface described in (ref. 26) and available at (ref. 28).

2.3.1. Forgery by obliteration. Different pre-processing techniques were tested: SNV (Standard Normal Variate), MSC (Multiplicative Signal Correction), Savitzky-Golay1st and 2nd derivative (2nd and 3rd order polynomial with window's width of 7, 9 and 15 points). PCA was performed and the score images were evaluated to identify the text hidden below the scrawl. MCR-ALS was also performed.

2.3.2. Forgery by adding text. The same pre-processing techniques as in the case of obliteration were evaluated for the adding text problem. Both PCA and MCR-ALS were carried out on the image spectra to identify whether more than one pen had been used to produce the document.

2.3.3. Sequence of intersecting lines. SNV, MSC, Savitzky-Golay 2nd derivative (3rd order polynomial and window widths of 7, 9 and 15 points) and autoscaling techniques were evaluated to pre-process the data set. MSC, using as the reference the spectrum of each pen, was also evaluated. This reference spectrum was obtained by taking an average spectrum from a sample region where only ink (on the paper) was present. An average spectrum was also obtained for the paper and toner. In addition, PLS-DA was performed with the pre-processed data set to select the main significant variables (*i.e.* wavelengths) that could be used to discriminate among the different components in the sample (paper, ink and toner). Thus, for each pair of samples (the ones produced with the same pen and same type of toner), a PLS-DA model was built and used for variable selection. Three regions of the sample containing only paper, only ink (on the paper) or only toner (on the paper) were used to build the model. Using the regression coefficients and the weights from the latent variables, the discriminant wavelengths were selected and used as input for the MCR-ALS algorithm. The concentration maps of each sample were obtained and analyzed to identify the sequence of the events used to produce the intersecting lines, *i.e.* to identify if the pen ink lay above the toner line or below, which could imply a possible forgery.

3. Results and discussion

3.1. Forgery by obliteration

From all the pre-processing techniques evaluated to preprocess the data set, the best results were obtained using SNV. After the preprocessing stage, PCA was performed with each image and the first 10 principal components were calculated. Fig. 3 shows the score images from three different samples. The first column (Fig. 3a) shows the score images of the two first PCs related to Brand 2 of the gel pens (G2). It is possible to distinguish clearly the text that has been covered by one ballpoint pen (BP5). The PC1 exhibits the variability between the G2 ink spectrum and the paper spectrum, while PC2 explains the variability between the G2 and BP5 spectra. Thus, it is possible to visualize the text that was hidden and the adulteration made with the other pen. In the second column (Fig. 3b), it is also possible to see the text in the PC2 score image, showing that PC2 explains the differences between the ink used to write the text (BP2) and the paper. PC1, on the other hand, shows the variability between the spectrum of the ink used to hide the text (BP3) and the paper spectrum. Fig. 3c illustrates the results of an obliterated sample where PCA could not differentiate the spectrum of the ink used to write the text (BP1) from the paper or from the second pen (G2).

Table 1 summarizes the results obtained for obliterated samples. The squares with a dash (-) represent those cases where it was not possible to see clearly the hidden text. The squares with an "OK" label represent the cases where it was possible to see the text in, at least, one out of the 10 PCs. The black squares show the combinations that were not used (same pen).

Thirty nine (39) texts out of ninety (90) samples were successfully identified. These correspond to texts written using pens BP5 and G2, regardless of the pen used to hide the text. Results were also successful with pens BP2, BP4 and G1. When the three brands of ballpoint pen (BP1, BP3 and BP6), rollerball (RB1) and hydrographic (H1) pens were used to write the text, it was not possible to identify it in any of the 10 calculated PCs.

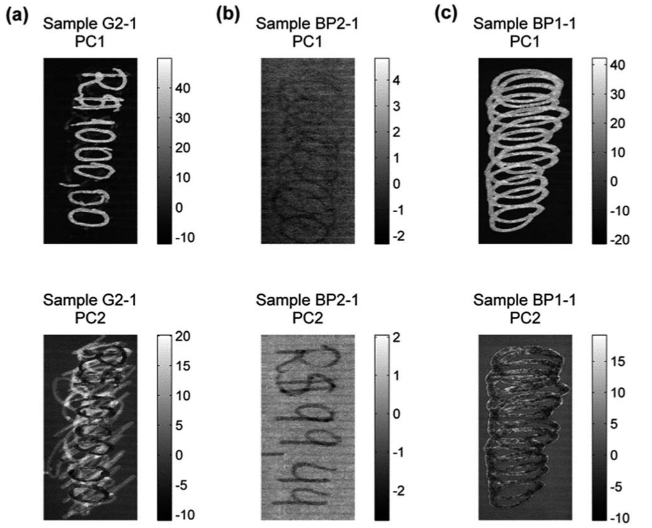

Fig. 3 Images of three samples produced by the scores of the two first PCs: (a) G2-1: BP5 was used to hide the text produced with G2; (b) BP2-1: BP2 was used to write the text that was hidden with BP3; (c) BP1-1: BP1 was used to write the text and G2 used to hide it.

Table 1 Summary of results obtained for obliterated samples

		Pen used to hide the text									
	Brand	BP1	BP2	BP3	BP4	BP5	BP6	G1	G2	RB1	H1
Pen used to write the text	BP1	■	-	-	-	-	-	-	-	-	-
	BP2	Ok	■	Ok	Ok	Ok	Ok	Ok	-	Ok	Ok
	BP3	-	-	■	-	-	-	-	-	-	-
	BP4	Ok	Ok	Ok	■	-	Ok	Ok	-	Ok	Ok
	BP5	Ok	Ok	Ok	Ok	■	Ok	Ok	Ok	Ok	Ok
	BP6	-	-	-	-	-	■	-	-	-	-
	G1	Ok	-	Ok	Ok	-	Ok	■	-	Ok	Ok
	G2	Ok	Ok	Ok	Ok	Ok	Ok	Ok	■	Ok	Ok
	Rb1	-	-	-	-	-	-	-	-	■	-
	H1	-	-	-	-	-	-	-	-	-	■

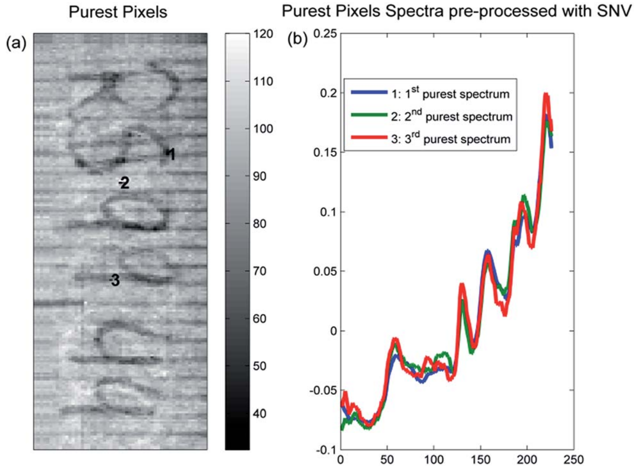

Fig. 4 (a) Purest pixels location; (b) purest pixels spectra pre-processed with SNV and used as initial estimates in MCR-ALS.

MCR-ALS was also employed to improve the PCA results. Fig. 4 shows the purest pixel spectra found in one of the samples (Fig. 2) and the respective pixels in the sample image. It is possible to see that the algorithm was capable of identifying the three most different compounds in the sample BP-1 used as an example (Fig. 3b). The pure spectra were used as input for the MCR-ALS algorithm and the distribution maps for each pure spectrum are shown in Fig. 5.

It is possible to notice that the purest pixel selected, represented in Fig. 5a, is related to the information regarding the ink used to write the text because the white pixels represent the high relative concentration of the respective compound. The second distribution map (Fig. 5b) corresponds to the ink of the pen used to hide the text, while the third distribution map (Fig. 5c) is associated with the spectral information of the paper.

The MCR-ALS approach successfully identified the hidden text in the same samples as PCA. The samples that could not be identified by PCA, could not be identified by MCR-ALS as well.

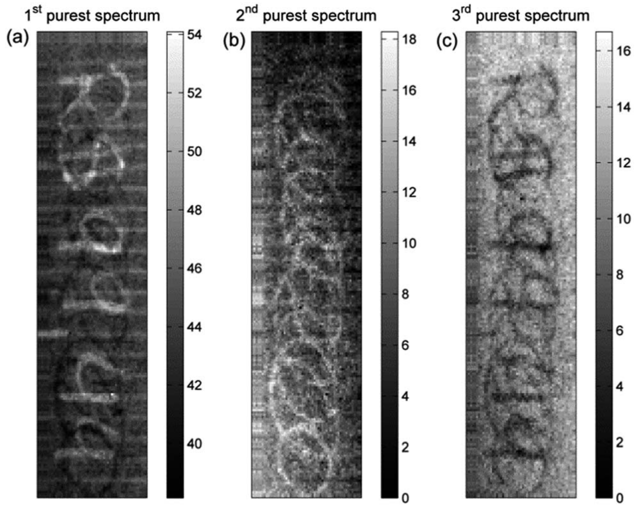

Fig. 5 Distribution maps for the first three purest pixel estimations of MCR-ALS.

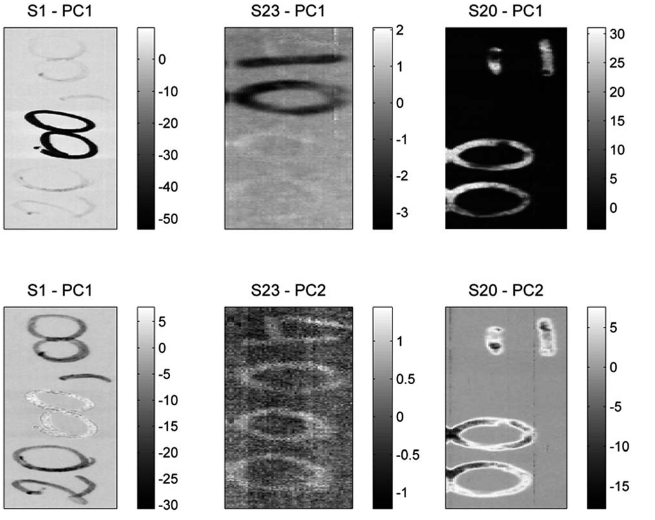

Fig. 6 Score images of the principal component analysis from three different samples. PC1 and PC2 of Samples S1, S23 and S20 are shown in Fig. 2d–f, respectively.

Hence, in this case of forgery, PCA is recommended because it is simpler and has the same potential as MCR-ALS.

3.2. Falsification by adding text

Among all the pre-processing techniques used to treat the data set, SNV showed the best results. After preprocessing, a PCA model was built and the first 10 PCs were obtained. Fig. 6 shows the score images related to the two PCs of three different samples, used as an example. As commented in Section 2, sample S1 was produced by adding two zeroes to the number 20 to form 2000. In sample S23, the number 10 was transformed into 4000 by modifying the number 1 into a 4 and adding two zeroes. In sample S20, the number 10 was also modified and transformed into 7000, as shown in Fig. 2d–f.

The first two PCs from sample S1, which combined the pens G2 and BP5 show that PC1 explains the variability of G2 (the pen used to forge the initial text), while PC2 explains the contrast between BP5 and G2. The analyst is able to see clearly that the document was produced with more than one pen.

The information captured by PC1 in Sample S23 is highly related to BP2 and paper spectra, but also shows information related to BP8 (light grey pixels). However, PC2 does not show any difference between the two pens used; rather, it shows the variability between the spectral information of the paper and the inks. The 1st and 2nd PCs of sample S20 show the variability of paper and BP9 spectra. It is not possible to identify the information related to the other pen used to simulate the forgery (BP1), even though the analyst could compare it with the original sample and certify that different pens were used to produce the document in question.

In 4 out of 22 samples, it was not possible to discriminate among the inks using the score images. In fact, in those cases, it was not possible to differentiate the inks from the paper. Nevertheless, 82% of samples were successfully identified. MCR-ALS was also carried out but did not show any improvement.

3.3. Sequence of intersecting lines

Unsatisfactory results were obtained from the dataset pre-processed with SNV and MSC using the average spectra as reference and autoscaling techniques. This is probably because the paper and the ink spectra were very similar. In addition, the toner spectrum was very different from that of the paper and ink. Thus, any preprocessing technique that uses average spectrum from the entire spectra sample set would provide a loss in information regarding the ink because the toner spectrum has considerably higher reflectance values in the NIR range. The best results were obtained with MSC employing the spectrum of the pen ink as reference, as described in the Experimental section. The proper selection of the reference spectrum is important to enhance the ink information on the data set to make it distinct from the paper spectrum.

The dataset, after preprocessing with MSC, was submitted to PLS-DA. Because the direct classification carried out by PLS-DA did not show a regular pattern to identify the sequence of the line at the intersecting point, this technique was used as a variable selection tool, as described below. To perform the PLS-DA, three different regions, squares with 20 × 10 pixels, of the sample containing only paper, ink or toner, were used to build a model with 4 latent variables.

The prediction of the training set of each pen was analyzed. Fig. 7 shows this information for two different samples produced with different types of pens. Note that the toner is predicted with high accuracy. However, some pen inks are confused with the paper, as shown in Fig. 7 (sample 2.1).

The weights and regression coefficients from the PLS-DA model were evaluated and the most important information for all the samples was provided by the 2nd and 3rd latent variables (see Fig. 8b and c for the results of sample 1.1). The first latent variable was related to the mean spectrum because data were not mean-centered (Fig. 8a). The wavelengths were selected by observing their weights: the ones that showed the highest and the lowest weight values were selected to evaluate the regression coefficients to guarantee that uninformative variables were left out. Different variables were selected for each pair of samples.

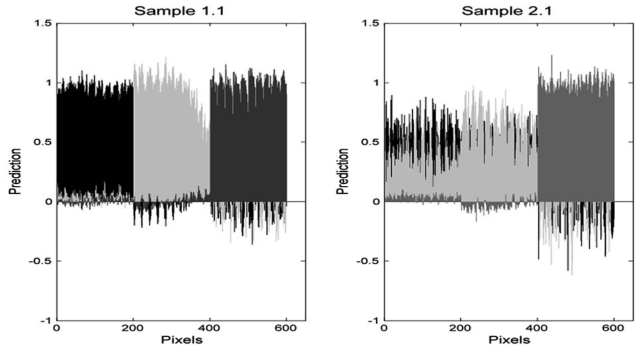

Fig. 7 Training set prediction from 2 different samples with different types of pens. The black bars (pixels from 0–200) refer to pixels containing only paper information; the light grey bars (pixels from 201–400) refer to pixels with information from ink over paper and the dark grey bars (pixels from 401–600), refer to pixels with information from toner over paper.

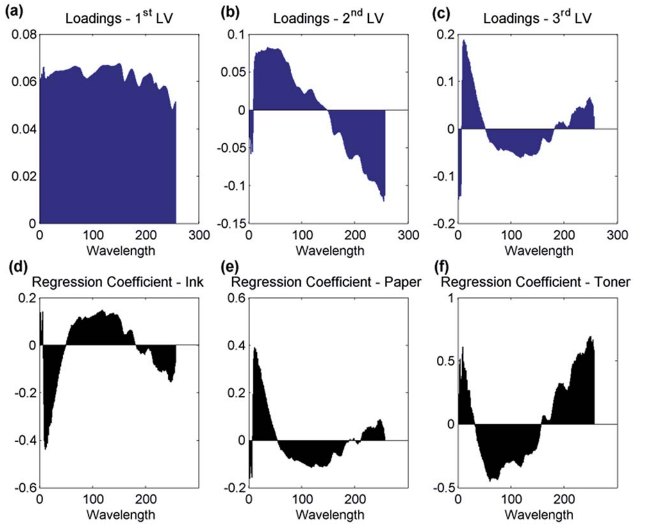

Fig. 8 Sample 1.1: loading plots from the PLS-DA model for the first three latent variables (a), (b) and (c); regression coefficient plots of PLS-DA model for (d) ink, (e) paper and (f) toner.

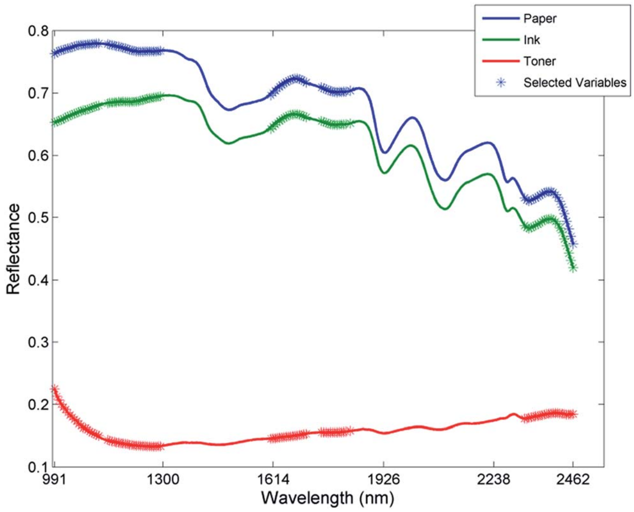

Fig. 9 PLS-DA selected variables (scatter dots) from paper (blue line), ink (green line) and toner (red line).

Fig. 8d–f show the regression coefficients for the PLS-DA model, emphasizing the most relevant spectral ranges. Fig. 9 shows the variables selected by PLS-DA. The results obtained were similar to other samples analyzed, which imparts confidence to the selection carried out.

The reflectance values of the selected wavelengths were used as input to the MCR-ALS, and non-negativity constraints were imposed on the concentration and the spectral direction. The distribution maps of each compound (paper, ink and toner) were obtained and studied for each sample. Fig. 10 shows those maps for two pairs of samples: samples 1.1 and 1.2, and samples 5.1 and 5.2. Samples 1.1 and 1.2 were produced with the same pen, but in 1.1 the ink line lies over the toner while in 1.2 the ink line lies below the toner (the same order for samples 5.1 and 5.2). The red pixels represent high values of the compound and the blue pixels represent low values. Three figures in the first row show the distribution map from paper (left), ink (center) and toner (right) obtained after the MCR-ALS for sample 1.1. It is possible to notice that distribution maps, showed in Fig. 10, can easily differentiate between the compounds in the samples. When the ink line lies under the toner line, it is possible to see a gap at the intersection point of the toner distribution map (see toner distribution map for samples 1.2 and 5.2 in Fig. 10). When the ink line lies above the toner line, this gap is not observed (see toner distribution map for samples 1.1 and 5.1 in Fig. 10). From the 20 pairs of evaluated samples, 17 exhibited the same behavior.

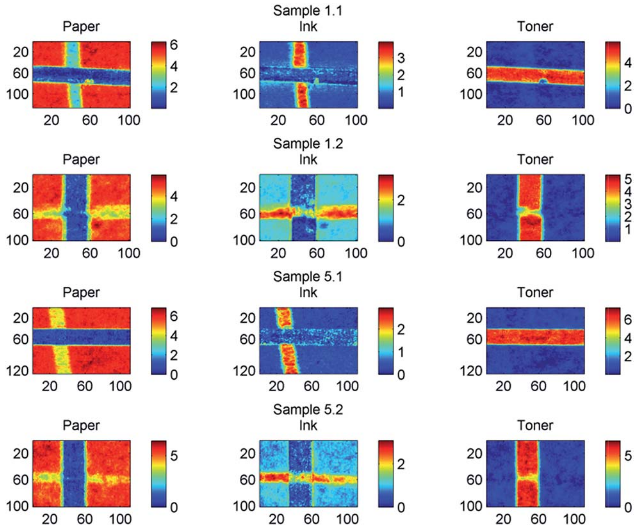

Fig. 10 Distribution maps for two different pairs of samples. The 1st, 2nd and 3rd columns show the paper, ink and toner distribution maps, respectively, for samples 1.1 (first row), 1.2 (second row), 5.1 (third row) and 5.2 (fourth row). In samples 1.1 and 5.1 the ink line lies over the toner line; in samples 1.2 and 5.2 the ink line lies below the toner.

Ozbek *et al.* analyzed the sequence of writing/printing instruments used to produce samples.[29] Inks from different printers (inkjet and toner), blue and red pens were used to produce samples of superimposed layers of ink. A transversal cut was made into the paper to analyze the layers with a microscope. The authors reported the difficulty of analyzing samples produced with dark inks and inks with the same color. As described by the authors, a laser printer used powdered ink particles that were electrically charged by a drum, heated and finally fused onto paper. Because the ink is fused, there was little penetration into the paper; instead, the ink cooled and bound to the cellulose fibers. They also verified that when the toner layer was on the top of the pen ink layer, a double layer formed. When the ink pen was on the top of the toner, the pen ink showed a tendency to penetrate through the toner ink, leading to ink mixtures.

The abovementioned conclusions are in accordance with the present results. The gap observed on the toner line when the ink line lies under the toner line might be because of the fact that the paper surface is very important for toner deposit. Then,

Table 2 Summary of results obtained for intersecting lines samples

Type	Brand	Toner 1	Toner 2
Ballpoint pen (BP)	1	OK	OK
	2	OK	OK
	3	OK	OK
	4	OK	OK
	5	OK	OK
	6	OK	OK
Hydrographic (H)	1	X	X
Gel (G)	1	OK	OK
	2	OK	X
Rollerball (RB)	1	OK	OK

when the toner is placed on a surface previously modified by the presence of ink, the amount and distribution of the toner is likely to be different, and a gap appears at the toner line. The toner shows high absorption through most of the NIR range, therefore, when the toner line lies below the ink, it is not possible to observe the gap because the toner reflectance spectrum dominates the ink spectrum, and the latter cannot be distinguished.

Table 2 summarizes the results obtained for each pair of samples analyzed. All the pairs could be successfully identified whenever ballpoint, roller ball, and one of the two brands of gel pens were used. Only the hydrographic pen and one of the two brands of gel pens showed problems for identification.

In the case of the hydrographic pen (using the two different types of toner), it was not possible to detect the ink in the image using the techniques chosen or to see the gap at the intersection line. In the case of the pair of samples prepared with one brand of gel pen (G2) and type of Toner 2, the gap appeared in both the samples (see Table 2). It is important to note that this gel pen shows a very particular spectrum, with high reflectance values in the near infrared range. Nevertheless, most of the samples, 17 pairs out of 20, were successfully identified. For forensic purposes, the analysis can be very helpful because the presence of a gap at the toner line in the distribution map of the toner suggests that the toner was placed after the ink, indicating a possible forgery.

In general, it possible to notice that the ink spectrum is highly superimposed by the paper spectrum because cellulose shows high absorbance values in the NIR range. The extreme case is the brand of hydrographic pen (H1), where its spectra are hardly differentiated from the paper spectra, causing the results of types (a) and (c) to be unsuccessful (when used to write the text). On the other hand, the chemical composition of one brand of gel pen (G2) provided significantly different spectra from paper, showing high reflectance in the near infrared range. This pen allows an easy identification of forgery by obliteration (type a) and adding text (type b) when this pen is used to write the text. However, in the case of intersecting lines, the behavior of this pen could be difficult to predict.

4. Conclusions

In this work, a chemometric–based methodology for forgery identification supported by HSI-NIR imaging has been provided. Different models and preprocessing techniques have been applied and compared depending on the type of problem approached. For type (a), obliteration, the hidden texts of 39 out of 90 samples (43%) were identified by PCA and MCR-ALS. For forgery by adding text (type b), at least one out of two pens used to produce the samples were identified in 82% of the samples (18 out of 22). Finally for the crossing lines forgery (type c), the presence of a gap at the toner distribution map was used as a criterion to identify if the ink line was made before printing toner. This was successfully identified in 85% of the samples (17 out of 20). Although not all samples were successfully identified, the HSI-NIR associated with chemometric tools shows excellent potential for detecting document forgery. The methods proposed are more objective and less dependent on personal judgments as the traditional ones are usually based in visual inspection by a very skilled person.

Acknowledgements

INCTAA (Processes no. : CNPq 573894/2008-6; FAPESP 2008/57808-1), NUQAAPE, FACEPE, CNPq, CAPES, Spanish Ministry of Science and Innovation – MICINN (grant DPI2011-28112-C04-02).

Notes and references

1 C. D. Adam, S. L. Sherratt and V. L. Zholobenko, *Forensic Sci. Int.*, 2008, **174**, 16–25.

2 N. C. Thanasoulias, N. A. Parisis and N. P. Evmiridis, *Forensic Sci. Int.*, 2003, **138**, 75–84.

3 J. Zieba-Palus and M. Kunicki, *Forensic Sci. Int.*, 2006, **158**, 164–172.

4 A. Kher, M. Mulholland, E. Green and B. Reedy, *Vib. Spectrosc.*, 2006, **40**, 270–277.

5 L. K. Ng, P. Lafontaine and L. Brazeau, *J. Forensic Sci.*, 2002, **47**, 1238–1247.

6 J. Zlotnick, *Forensic Sci. Int.*, 1998, **92**, 269–280.

7 R. L. Brunelle and K. R. Crawford, in *Advances in the Forensic Analysis and Dating of Writing Ink*, ed. C. Charles, Thomas Pub Ltd, Springfield, 2003, ch. 2, pp. 9–12.

8 P. M. Lalli, G. B. Sanvido, J. S. Garcia, R. Haddad, R. G. Cosso, D. R. J. Maia, J. J. Zacca, A. O. Maldaner and M. N. Eberlin, *Analyst*, 2010, **135**, 745–750.

9 G. Reed, K. Savage, D. Edwards and N. Nic Daeid, *Sci. Justice*, 2014, **54**, 71–80.

10 M. Tahtouh, P. Despland, R. Shimmon, J. R. Kalman and B. J. Reedy, *J. Forensic Sci.*, 2007, **52**, 1089–1096.

11 T. Chen, Z. D. Schultz and I. W. Levin, *Analyst*, 2009, **134**, 1902–1904.

12 G. Edelman, T. G. Van Leeuwen and M. C. G. Aalders, *Forensic Sci. Int.*, 2012, **223**, 72–77.

13 R. L. Schuler, P. E. Kish and C. A. Plese, *J. Forensic Sci.*, 2012, **57**, 1562–1569.

14 A. Chaikovsky, S. Brown, L. S. David, A. Balman and A. Barzovski, Color, *J. Forensic Sci.*, 2003, **48**, 1396–1405.

15 R. Kaur, K. Saini and N. C. Sood, *Sci. Justice*, 2013, **53**, 206–211.

16 C. E. H. Berger, *Sci. Justice*, 2013, **3**, 55–59.
17 K. Bojko, C. Roux and B. J. Reedy, *J. Forensic Sci.*, 2008, **53**, 1458–1467.
18 G. Payne, C. Wallace, B. Reedy, C. Lennard, R. Schuler, D. Exline and C. Roux, *Talanta*, 2005, **67**, 334–344.
19 A. Widener and C. Drahl, *Chem. Eng. News*, 2014, **92**(19), 10–15.
20 P. Geladi, H. F. Grahn and J. E. Burger, in *Techniques and Applications of Hyperspectral Image Analysis*, ed. P. Geladi and H. F. Grahn, John Wiley & Sons Ltd, West Sussex, 2007, ch. 1, pp. 1–15.
21 S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37–52.
22 P. Geladi, H. Isaksson, L. Lindqvist, S. Wold and K. Esbensen, *Chemom. Intell. Lab. Syst.*, 1989, **5**, 209–220.
23 J. M. Prats-Montalbán, A. De Juan and A. Ferrer, *Chemom. Intell. Lab. Syst.*, 2011, **107**, 1–23.
24 A. De Juan, M. Maeder, T. Hancewicz, L. Duponchel and R. Tauler, in *Infrared and Raman Spectroscopic Imaging*, ed. R. Salzer and H. W. Siesler, WILEY-VCH, Weinheim, 2009, ch. 2, pp. 65–106.
25 R. Tauler, *Chemom. Intell. Lab. Syst.*, 1995, **30**, 133–146.
26 J. Jaumot, R. Gargallo, A. De Juan and R. Tauler, *Chemom. Intell. Lab. Syst.*, 2005, **76**, 101–110.
27 W. Windig and J. Guilment, *Anal. Chem.*, 1991, **63**, 1425–1432.
28 Multivariate Curve Resolution Homepage, http://www.mcrals.info/.
29 N. Ozbek, A. Braz, M. López-López and C. García-Ruiz, *Forensic Sci. Int.*, 2014, **234**, 39–44.