Pre-processing of inkjet prints NIR spectral data for principal component analysis

Michal Oravec, Lukáš Gál, Michal Čeppan

Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, SK-812 37 Bratislava, Slovak Republic
michal.oravec@stuba.sk

Abstract: This paper presents a novel approach in non-destructive analysis of inkjet-printed documents. Our method is based on the combination of molecular spectroscopy in the Near Infrared Region (NIR) and a chemometric method — principal component analysis (PCA). The aim of this work was to prepare spectral data for the analysis of the interrelationships between 19 samples consisting of the same type of office paper on which black squares were full printed in black ink only. The spectra were obtained separately using the Ocean Optics System in two spectral regions, i.e., overtones: 1000—1600 nm and combination bands: 1600—2300 nm, with the paper base. Experimental results confirmed the high reliability of the proposed approach despite the sparse dataset.

Keywords: NIR spectra pre-processing, principal component analysis

Introduction

There are a large number of printed documents that have to be investigated in forensic analysis. It is crucial to reveal the origin of the printed documents. These documents can be printed using a large number of different printing devices. In general, we distinguish between *inkjet* and *laser* printers. Moreover, the task is to find the origin of a printed document when the cartridge was not produced by an official manufacturer. Chemical composition of these inks is usually different than the original one and the unoriginal cartridges are highly penetrated as they are produced at significantly lower costs (Tchan, 2007).

Basic investigation of the graphic document is focused on its material analysis. The most frequently used analytical techniques: TLC, HPTLC, GC-MS, and HPLC (Wilson et al., 2004), (Weyermann et al., 2007), (Hofer, 2004) are based on destructive preprocessing of the document samples. Application of these methods leads to serious destruction of the analyzed document, which is the major drawback of these methods (Gál et al., 2014).

Nine samples of inkjet prints were analyzed by Raman and SERS methods by Braz et al. (2014). The advantage of the Raman method is that the printed document is not destructed. However, sparse signals of the black ink in the samples were obtained due to the fluorescence effect. On the other hand, the slightly destructive method SERS leads to the significantly more pronounced results of the black ink samples.

The investigated documents are complex, because they consist of the paper-base and the material structure of the graphic information. The main goal of this study was to minimize the destruction of the inkjet-printed document. The well-known molecular spectroscopy method represents a suitable non-destructive approach. Usually, the obtained results are not clear, as they contain also some measurement noise as well as undesirable information about the background (Cen et al., 2007).

Chemometric methods play an important role in sample analysis as they are fast and effective, enabling evaluation of the spectra in correlation to the physical, chemical, and analytical characteristics of the examined sample (McClure, 2003). Successful application of the chemometric methods requires some degree of experiences with these methods (Lavine, 1998). The well-known principal component analysis (PCA) is a widely-used method for effective identification of the relationship between the samples. This strategy was successfully implemented in a wide range of mathematical tools (Jolliffe, 2002) as a suitable method for large data sets. The main advantage of PCA is the minimum loss of useful data (Meloun et al., 2005). The essence of the method is to find optimal position of the first principal component, the vector in the space determined by the samples. The second principal component is located orthogonally to the first one and defines the second highest value of the total variance. The process proceeds until the all total variance is completely described by principal components (Meloun et al., 2005; Esbensen, 2002; Marida et al., 1979).

Near Infrared Region (NIR) analysis is fast, but the interpretation of the obtained data is difficult.

Our aim was to develop a method generating accurate and stable data for the purpose of forensic analysis. Hence, the task was to find an optimal

combination of transformation methods; particularly to prepare spectral data for the investigation of the interrelationships between 19 different samples. The measured data were pre-processed for the PCA analysis classifying the samples considering their properties. The obtained groups of samples were generated with regard to the similar shape of their spectra.

Experimental

In this study, 19 samples of the inkjet-printed documents were analyzed and their characteristics are summarized in Tab. 1.
For each sample, a list of chemical compounds was prepared, see Tab. 2. These data were collected based on the Material Safety Data Sheet (MSDS online). However, MSDS of non-genuine inks and two Epson samples are not available, yet. Data highlighted in red color contain a chemical compound called *Carbon black*.

For the analysis, inkjet prints from three manufacturers (HP, Epson and Canon) were used. The samples consist of the office paper (Xerox Performer 80g/m^2) on which black squares were full printed in black ink only. The spectra were pre-processed and analyzed by PCA in two spectral regions, i.e., overtones: 1000–1600 nm and combination bands: 1600–2300 nm, separately. These mathematical operations were realized using the software *The Unscrambler X*. The fundamental part

Tab. 1. Inkjet printers used for sample preparation.

Sample	Manufacturer	Type of Print Devices
C1	Canon	IP 3000
C2	Canon	Pro 9500 II
C3	Canon	IP 4300
E1	EPSON	SX 425
E2	EPSON	PMD 800
E3	EPSON	PM 830c
E4	EPSON	PX 730WD – origin. ink
E5	EPSON	DX 7400
E6	EPSON	PX 730 WD – unoriginal ink
E7	EPSON	P50 – unoriginal ink
E8	EPSON	P50 – origin. Ink
E9	EPSON	SX 130 – unoriginal ink
E10	EPSON	L 210
H1	HP	Photosmart C4580
H2	HP	Photosmart 3210
H3	HP	Photosmart C3180
H4	HP	DeskJet 920C
H5	HP	Photosmart C1410
H6	HP	Office jet 8100

Tab. 2. Chemical compounds mentioned in MSDS.

Sample:	C1	C2	C3	E1	E4	E5	E7	E10	H1	H2	H3	H4	H5	H6
1,5-pentanediol									X		X			
2-pyrrolidone									X	X	X	X	X	X
Ammonium benzoate	X													
Carbon Black				X		X			X		X	X	X	X
Colorants					X		X	X						
Cyclo amid														X
Diethylene glycol	X	X												
Etylene glycol			X											
Glycerin	X	X	X							X				
Glycerols				X	X	X	X	X						
Heterocyclic compound			X											
Isopropyl alcohol												X	X	
Lactam		X												
Proprietary organic materials				X	X	X	X	X						
Substituted diol									X					
Substituted naphthalene sulfonic acid			X											
TEGBE*				X	X	X	X	X						
Triol	X							X						
Water	X	X	X	X	X	X	X	X	X		X	X	X	X

*trietylene glycol butyl ether.

of this work was to design an appropriate procedure of raw spectral data pre-processing for their analysis by PCA.

Results and discussion

The combination of transformation methods was selected with regard to optimal results obtained by PCA models; therefore, our aim was to achieve the maximum possible value of variance described by the first two principal components (Oravec et al., 2014). The pre-processing transformation methods for NIR spectra were chosen as follows: interpolation (Fig. 1), smoothing filter Savitzky–Golay (Fig. 2), Standard Normal Variate (SNV) (Fig. 3), detrending (Fig. 4), baseline correction (Fig. 5), and normalization – Mean (Fig. 6). These transformation methods were applied to all investigated samples. The wavelength measured in nanometres can be found from x-axes in Figs. 1–6. The y-axes in these figures show the values of optical density *D*, or those of normalized optical density.

Interpolated spectra of 19 samples were divided into two groups, see Fig. 1. Sample C3 belongs to the group of spectra with the density value around one (red lines in Fig. 1). Although, based on MSDS, samples C1, C2, and C3 do not include the component *Carbon black*, they also belong to the above-mentioned group (red lines in Fig. 1). These spectra were measured by NIR spectroscopy. Preliminary PCA of interpolated spectra only showed 100 % total variance of the first principal component.

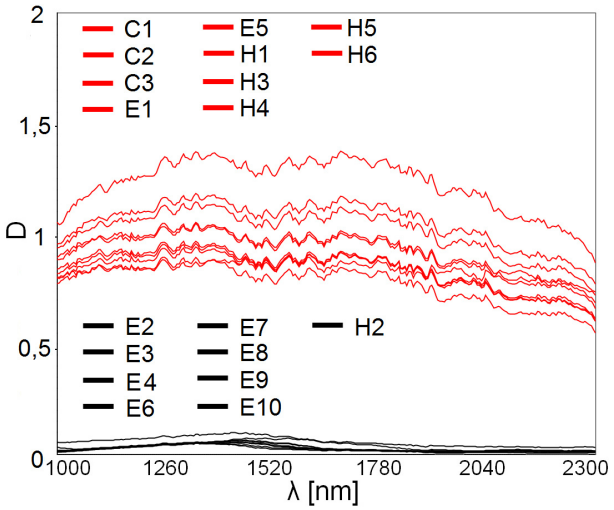

Fig. 1. Interpolation in the range from 1000 to 2300 nm.

Using the Savitzky–Golay smoothing method, the influence of random noise in the spectra was reduced (Fig. 2). The principle of this method is that data measured in small wavelength intervals can be recalculated and replaced by a function of the appropriate degree of the polynomial. Data pre-processed by the Savitzky–Golay method provide a better estimation in comparison with the raw unprocessed spectra (Savitzky et al., 1964).

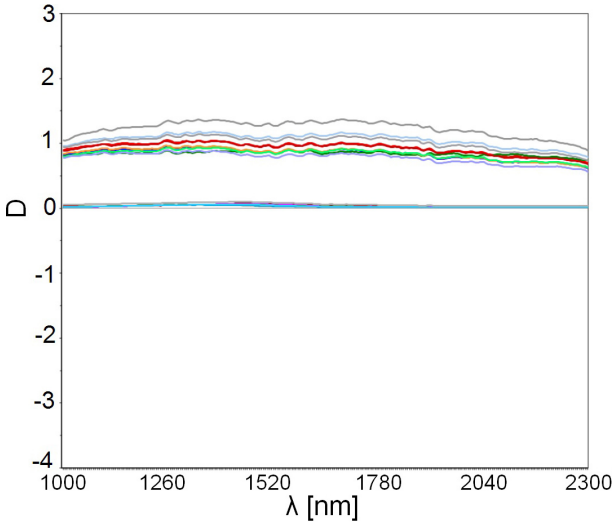

Fig. 2. Application of the Savitzky–Golay smoothing filter.

To suppress the heterogeneity and anisotropy effects of the samples surface, mathematical transformation based on the Standard Normal Variate (SNV) and detrending were applied to the spectra (Fig. 3).

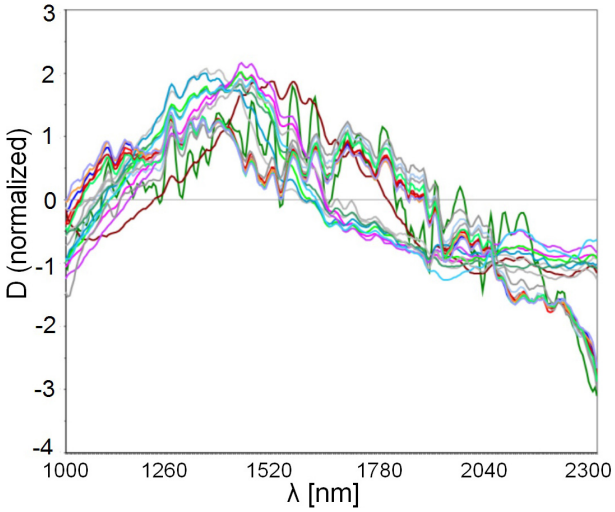

Fig. 3. SNV method implemented on smoothed spectra.

Detrending correction is used after the SNV method had been applied. This transformation method was applied to eliminate the trends in spectroscopic

data as it corrects the trends of all spectra around the zero value on the y-axis (Fig. 4). Using the combination of SNV and detrending resulted in corrected properties of multicollinearity, shifts, and baseline curvature (Candolfi et al., 1999; Barnes et al., 1989).

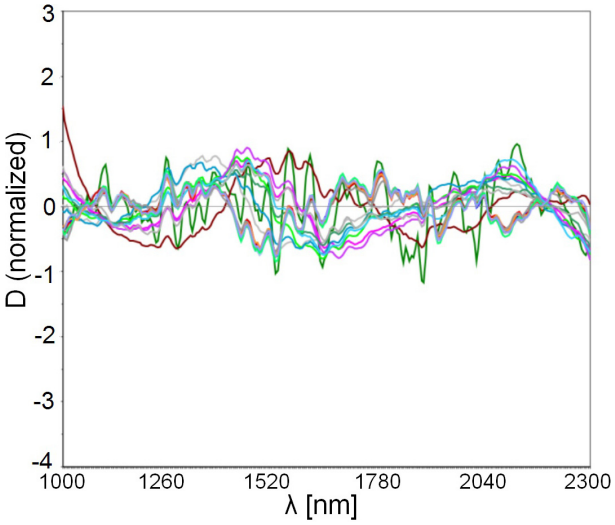

Fig. 4. Detrending transformation.

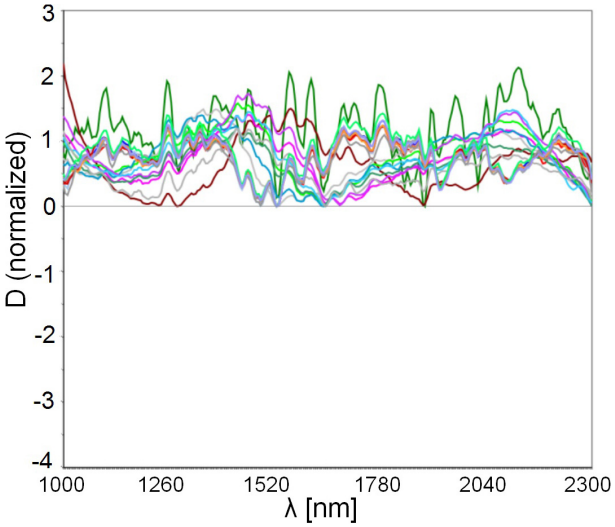

Fig. 5. Baseline correction.

Due to the baseline correction, minimum values of spectra are localized at the zero value of the y-axis (Fig. 5).

The mean normalization method divides every point of the spectrum by their average value (Fig. 6). NIR spectra prepared using this procedure are suitable for the PCA analysis.

The most important PCA outcomes are the scatter plots of the component scores, where the samples are assigned to groups based on their differences considering the pre-processed spectra of the samples in the two dimensional space. This strategy significantly reduces the human factor of samples sorting (Oravec et al., 2015; Gál et al., 2014). Fig. 7 shows the dependence of the explained variance on principal components. As can be seen, more than 95 % of explained variance is expressed by PC-1 and PC-2. To find an appropriate combination of transformation methods and their parameters, feedback is needed; results of the scatter plot and of the explained variance. Variance values for the first two principal components were explained using the PCA model in the range of 1000–1600 nm (Fig. 7); this value was higher compared to that obtained in the range of 1600–2300 nm (Fig. 8). Specifically, differences between the values of variance in the two PCA models were about 16.58 percentage points.

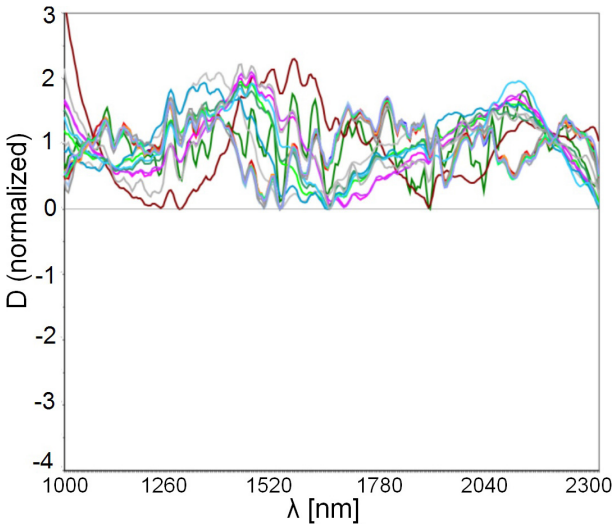

Fig. 6. Normalization mean.

As it can be seen in Fig. 8, two principal components are sufficient for a successful analysis. The samples can be sorted into two groups G1 and G2 (Fig. 9). PC-1 describes 86 % and PC-2 10 % of the explained variance. Majority of the samples are concentrated near the direction of PC-1. Group G2 is more compact than G1 which contains the same nine samples as group G3 (Fig. 10).

In the PCA model in the range of 1600–2300 nm, G1 samples are more scattered than in the range of 1000–1600 nm, i.e., G1 samples do not form a separate group in the PCA model in the range of 1600–2300 nm. This can be caused by the higher explained variance of the PCA model in the range of 1000–1600 nm.

Experimental results confirmed that the presented approach is suitable for the analysis of inkjet-printed documents. The method combines two strategies: non-destructive molecular spectroscopy in the NIR region providing sample spectra, and chemometric PCA for the determination of the latent structure of sample relationships. Obviously, this approach is limi-

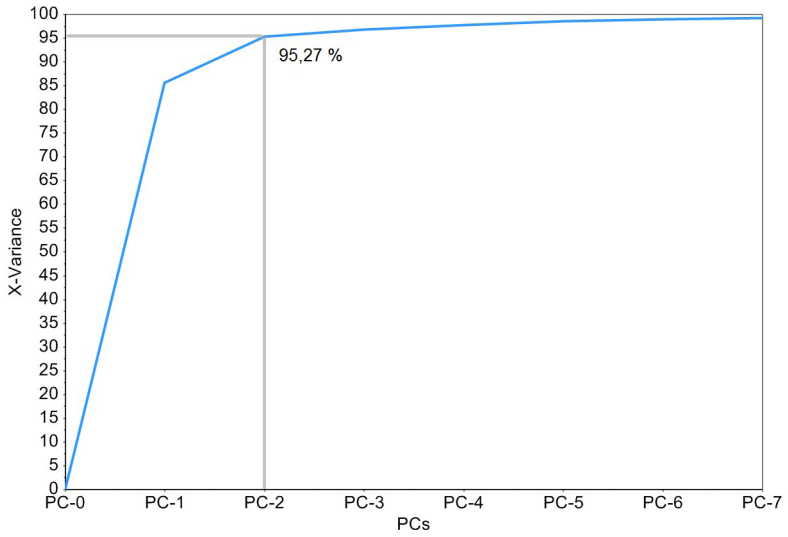

Fig. 7. Explained variance in the range of 1000–1600 nm with paper base.

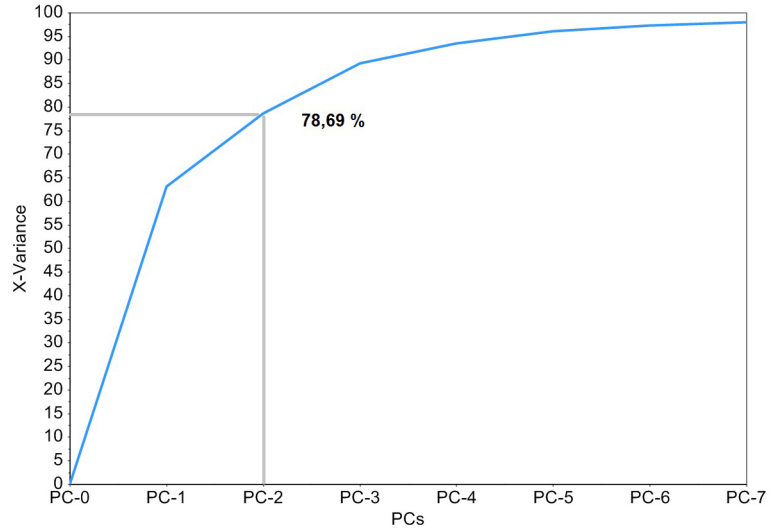

Fig. 8. Explained variance in the range of 1600–2300 nm with paper base.

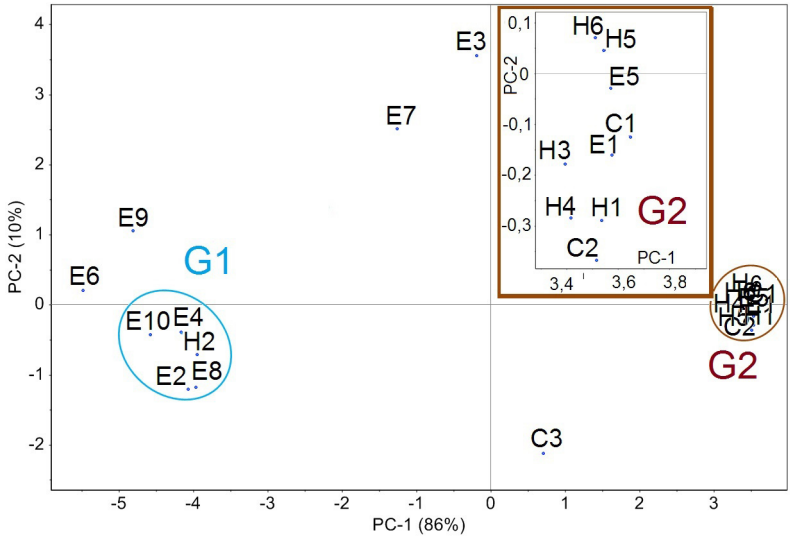

Fig. 9. PCA scores for printer ink spectra in the range of 1000–1600 nm.

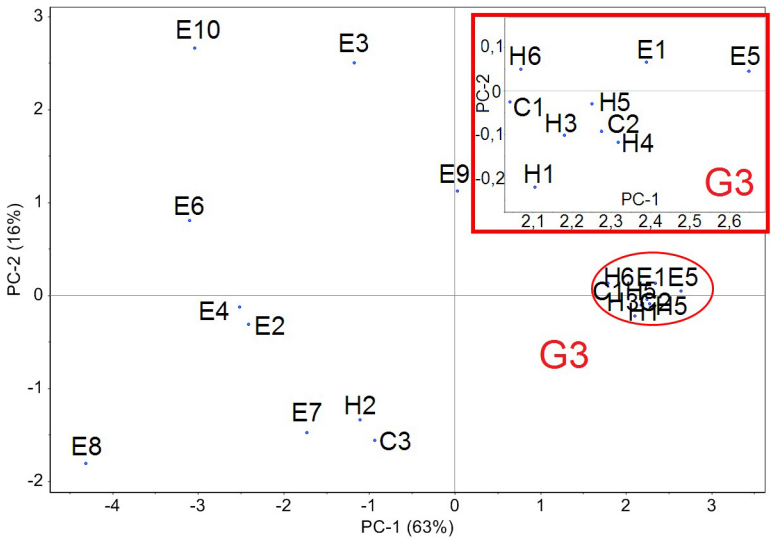

Fig. 10. PCA scores for printer ink spectra in the range of 1600–2300 nm.

ted by the insufficiently dense database. A database comprising 19 samples of three manufacturers has been generated. Experimental results confirmed the high reliability of the proposed approach although it is limited by the sparse dataset.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under the contract no. APVV-0324-10. This publication is the result of the project implementation: Centre of Excellence for Security Research ITMS code: 26240120034 supported by the Research & Development Operational Program funded by the ERDF. Michal Oravec was also provided an internal STU grant.

References

Barnes RJ, Dhanoa, MS, Lister SJ (1989) Appl. Spectrosc. 43: (5), 772–777.

Braz A, López-López M, Montalvo G, García Ruiz C (2014) Australian Journal of Forensic Sciences 47: 411–420.

Candolfi A, De Maesschalck R, Jouan-Rimbaud D, Hailey PA, Massart DL (1999) J. Pharm Biomed. Anal 21: 115–132.

Cen H, He Y (2007) Trends in Food Science & Technology 18: 72–83.

Esbensen K (2002) Multivariate Data Analysis – In Practice. 5., Oslo, CAMO Process AS, ISBN 82-993330-3-2.

Gál L, Oravec M, Čeppan M (2014) PCA and Fiber Optics VIS-NIR and NIR Reflectance Spectra for Examination of Inkjet Prints in Forensic Analysis, 7th International Symposium of Information and Graphic Arts Technology, Slovenia, June 5–6.

Hofer R (2004) Journal of Forensic Science 49: 1353–1357.

Jolliffe IT (2002) Principal Component Analysis, Second Edition. Springer-Verlag New York, ISBN 0-387-95442-2.

Lavine BK (1998) Anal. Chem 70: 209–228.

Marida KV, Kent JT, Bibby JM (1979) Academic Press Inc. 24: 502–1982.

McClure FW (2003) NIR Publications 11: 487–518.

Meloun M, Militký J, Hill M (2005) Computer analysis of multivariate data in examples (*in Czech*) Praha: Academia, ISBN 80-200-1335-0.

MSDS online. [accessible: 2015-10-30] http://www.ilpi.com/msds/.

Oravec M, Gál L (2014) Student Professional Conference Chemie is life in Brno, pp. 86, ISBN 978-80-214-5077-6.

Oravec M, Gál L, Čeppan M (2015) 16th Austrian Chemistry Days, University of Innsbruck.

Savitzky A, Golay MJE (1964) Anal. Chem 36: 1627–1639.

Tchan JS (2007) J. Imaging Sci. Technol 51: 299–309.

Wilson JD, La Porte GM, Cantu A (2004) Journal of Forensic Science 49: 364–370.

Weyermann C, Marquis R, Mazzella W (2007) Journal of Forensic Science 52: 216–220.