Appl. Phys. A (2016) 122:957
DOI 10.1007/s00339-016-0494-9

Applied Physics A
Materials Science & Processing

Hyperspectral imaging coupled with chemometric analysis for non-invasive differentiation of black pens

Damian K. Chlebda[1] · Alicja Majda[1] · Tomasz Łojewski[2] · Joanna Łojewska[1]

Received: 21 May 2016 / Accepted: 12 October 2016 / Published online: 18 October 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Differentiation of the written text can be performed with a non-invasive and non-contact tool that connects conventional imaging methods with spectroscopy. Hyperspectral imaging (HSI) is a relatively new and rapid analytical technique that can be applied in forensic science disciplines. It allows an image of the sample to be acquired, with full spectral information within every pixel. For this paper, HSI and three statistical methods (hierarchical cluster analysis, principal component analysis, and spectral angle mapper) were used to distinguish between traces of modern black gel pen inks. Non-invasiveness and high efficiency are among the unquestionable advantages of ink differentiation using HSI. It is also less time-consuming than traditional methods such as chromatography. In this study, a set of 45 modern gel pen ink marks deposited on a paper sheet were registered. The spectral characteristics embodied in every pixel were extracted from an image and analysed using statistical methods, externally and directly on the hypercube. As a result, different black gel inks deposited on paper can be distinguished and classified into several groups, in a non-invasive manner.

✉ Damian K. Chlebda
damian.chlebda@uj.edu.pl

[1] Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Cracow, Poland

[2] Faculty of Materials Science and Ceramics, AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Cracow, Poland

1 Introduction

Analysis of the written text is still a vexing issue. A number of papers cover the problem of written text analysis and differentiation. However, only few of them discuss the application of non-destructive instrumental methods, especially in relation to forensic and cultural heritage studies. Discrimination of different inks is essential in detecting forgery [1], but also seems to be critical in the study of historical documents. Analysis is mostly directed at backdating and chemical composition studies of writing media, and of assessing the state of preservation of documents [2, 3]. Nowadays, determination of ink formulas is widely performed using, for example, thin-layer chromatography of ink extracts [4] or capillary electrophoresis [5]. However, such methods are invasive and time-consuming [6], and these disadvantages must be considered. Different approaches, such as Raman spectroscopy [7–10], UV–Vis spectroscopy [4, 11], and IR spectroscopy [12], are based on the interaction of light with the substrate. Those methods ensure the physical integrity of the document, and yield data on organic and inorganic components of a sample. That allows for the identification of dyes and pigments used, and hence for ink differentiation.

Hyperspectral imaging (HSI), used in this study, is one of the modern spectroscopic techniques combining standard reflectance spectroscopy with photography, therefore enabling analysis of an object both ways. Spectral information collected from many channels (up to several hundred) enhances the capability of standard photography. HSI can be applied in different ranges of the electromagnetic spectrum. Depending on the sensor type, an image can be registered from ultraviolet (UV) to mid-infrared (MIR), or even to the far infrared range (FIR). Data are stored in a three-dimensional file called a hypercube or datacube. This

file has one spectral and two spatial dimensions [13]. Consequently, the information collected in the datacube allows materials to be identified and the registered image to be classified. HSI analysis, like photography, enables the contrast between selected features of an object and its surface to be enhanced. This kind of imaging therefore simplifies discrimination, especially when the sought-for features are invisible to the human eye (for example, text written in black ink on dark paper). An outstanding feature of this technique is that it allows non-destructive qualitative analysis, which is essential for research focused on fragile objects. HSI has been used successfully for the identification of areas with different compositions of covering materials, evaluation of object degradation, recovery of the underdrawings, and for the purpose of colour measurements [14, 15]. In forensic science, hyperspectral imaging has been used for the fingerprint testing, comparative analysis of blood and drugs traces, and analysis of fibres, with a view to determining their origin [16, 17]. Recently, and for the first time, HSI was successfully utilised in the analysis of micro-traces [16, 18] to evaluate the chemical changes taking place in different objects during the examination of a crime. As a non-destructive method of examination, HSI has also been applied to recognise fraud in documents [1]. Nevertheless, the differentiation issue requires further investigation.

Regarding the current demands in the fields of forensic and preservation science, hyperspectral imaging seems to be a suitable tool. A few papers discuss the use of HSI as an analytical technique for ink discrimination [19, 20]. The results depend strongly on the spectral properties of analysed inks. Comparison of inks performed only by detection of changes between different spectra does not always yield good and satisfying results, and enhancing standard investigation with multivariate statistical methods is necessary. Statistical methods have proved to be an efficient, time-saving, and more powerful tool than a classical visual comparison of samples or registered spectra [21, 22]. Differentiation of pen inks of the same colour and formed by mixtures of same dyes with additives cannot be accomplished due to high similarity of the spectra. Devices for automatic analysis of documents (for example, the Spectrum FORAM 685-2 by Foster and Freeman and the HSI Examiner 100 QD by ChemImage) [20] have become commercially available in recent years, but their relatively high price and limited technology make them unsuitable for use by every scientific institution. Still, optical analysis remains the cheapest approach for ink comparison and is often successful.

The aim of this work was to use HSI–VIS–NIR and multivariate analysis to identify a set of different kinds of common black gel pens produced by various manufacturers and available on the Polish market. The analyses were conducted to evaluate the possibility of identifying pens with inks of a similar colour, and to present the hyperspectral imaging technique as a non-destructive and non-invasive tool for document analysis. The literature about the forensic application of HSI mostly focuses on crime scene investigation, so we decided to investigate the problem of differentiation of writing traces by using HSI as an alternative to conventional approaches. The results of the application of HSI, presented further on in this paper, indicate that this technique might be used to distinguish between different gel ink pens of similar colour. The selection of an effective spectral range to 470–930 nm limited by camera sensitivity and intensity of the popular halogen light source is not conducive to easy visual differentiation; thus, we decided to apply a chemometric approach. This step included hierarchical cluster analysis (HCA), which is the basis for spectral classification and gives a general overview of the existing relationship in a dataset; principal component analysis (PCA), as a fast and robust method of spectral classification and analysis; and spectral angle mapper (SAM), an efficient and conceptually simple method that gives information about the spectral similarity of selected spectra. HCA and PCA have already proved to be powerful tools for discriminating between other types of writing media [7, 23], while SAM seems to be applied for this purpose for the first time. The combination of HSI with chemometric methods offers an easy methodology that does not require any expensive equipment and could be applied for routine analysis in the forensic laboratory.

2 Materials and methods

2.1 Sample characteristics

The study involved two independent sets of inks. The first is the test set of four selected black gel pen inks (Lexi 5, Bic Medium, Pentel BK437, and Paper Mate), which were applied on a paper support as three independent lines for each pen. This set was used to verify whether differences found between groups of inks were reproducible. The main analysis was performed on a set of 45 modern gel pen ink marks deposited on a paper sheet (see Fig. 1). A list of the 35 different pen inks tested (ten pairs were repeated) is given in Table 1. In order to verify the correctness of the classification, ten pairs of identical gel pen inks were chosen to validate further classification. The surveyed material was prepared by the Institute of Forensic Research (IES) in Cracow. All pens are produced by various manufacturers and available on the Polish market. These inks were chosen to reflect the colour most commonly encountered in casework.

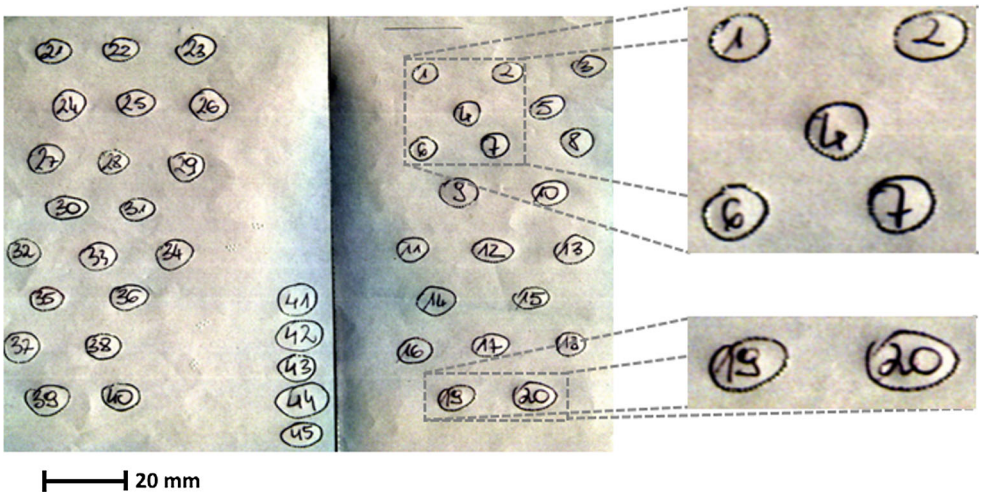


Fig. 1 Page with the 45 tested pen gel traces registered by hyperspectral imaging. The image presented is a result of a combination of different images derived from three representative wavelengths, namely 460 (*blue*), 550 (*green*), 640 (*red*) nm as RGB components

2.2 Hyperspectral imaging tests

Hyperspectral images were acquired using a hyperspectral camera (Headwall Photonics model VNIR C-series, Fitchburg, USA) in a push-broom configuration system (see Fig. 2). The camera was equipped with a C-Mount lens (Schneider–Kreuznach Xenoplan, F/1.4, FL 23 mm) with improved transmittance in the NIR range. The CCD (charge-coupled device) sensor used in the current study registered data in the spectral range from 369 to 1027 nm. The working distance (the distance between the document and the camera) was 60 cm. The final resolution of captured hyperspectral images of the primary dataset was 1392 × 1701 (pixels) × 658 nm (one image with a spatial dimension of 1392 × 1701 for every nanometre). Spatial resolution for this study (size of a pixel) equals 0.15 mm. The sample was illuminated by an optically stabilised halogen lamp (Fiberoptics SOL-R) back fitted with a Xenophot HLX 64635 halogen source. The line-light accessory was mounted at a distance of 30 cm from the surface of analysed document. The light beam was transferred through a line-light accessory equipped with a cylindrical lens. The heat received by the document from the light was not monitored. The camera was placed above the horizontal motorised stage, which enabled one axis movement of a sample. Data acquisition software was developed in the LabVIEW programming language (Laboratory Virtual Instrument Engineering Workbench) for self-use.

2.3 Data collection

The prepared sample with 45 traces of different black pen gel was scanned using the HSI system. The effective spectral range was 470–930 nm, due to low camera sensitivity and the low intensity of the light source beyond this range. The scan was performed in a stepwise mode with an average of 10 readings with 40 ms exposure time for each reading. This image registration path led to an improvement of the *S/N* ratio. During acquisition, the data were normalised on the fly using readings obtained for dark and white standards under the same conditions as used during scanning. Dark standard data (dark noise of the CCD camera) were acquired by covering the lens. A $BaSO_4$ plate was used as a white reference, giving the maximum reflectance value along the scanned line.

The light source variation and dark noise effect were corrected by applying the following equation:

$$I_n = \frac{I_{\text{sample}} - I_{\text{black}}}{I_{\text{white}} - I_{\text{black}}} \quad (1)$$

where I_n is the relative reflectance value and I_{sample}, I_{black}, I_{white} are the absolute reflectance values for the sample, the dark standard (∼0 % reflectance), and the white standard (∼99.9 % reflectance), respectively. As a result, a block of 658 images with dimensions of 1392 × 1701 was created. The 2D *xy* representation was extended in the third dimension *z* using spectral information collected for every single pixel. The hyperspectral image data were analysed using the ENVI 5.1 (Exelis Visual Information Solutions, Herndon, USA) and Spectronon Software (Resonon, USA). The extracted spectral information was further analysed using Statistica software version 12 (StatSoft Inc., Tulsa, USA). While the goal of using HSI is to obtain one spectrum per pixel, the spectral information contained in one pixel is influenced by external factors and acquisition systematic error. Thus, regions of interest (ROIs) have to be selected to extract average spectra for every ink line as representative spectra for every region. ROIs including at least 100 pixels were defined for each of the 45 traces of black pen gel. ROIs were selected in the areas with the highest chroma using

Table 1 List of gel pen brands used during analysis

No	Name	CIE $L^*a^*b^*$		
		L^*	a^*	b^*
1	Uni Lakubo Fine	22.53	1.41	−2.83
2	Uni Laknock Fine	20.46	4.67	−6.75
3	Uni Laknock II Fine SD-108	20.06	4.44	−6.79
4	NFI from TD	20.71	1.32	−2.67
5	BIC fine (orange)	21.62	3.91	−6.26
6	Renishaw	21.91	1.36	−2.48
7	BIC medium	19.03	2.06	−4.61
8	BIC N-S Fine	22.11	1.59	−3.15
9	Pilot BP-S Fine	19.40	1.74	−3.36
10	Pilot BPS-GP hFi	19.58	1.81	−3.38
11	Pentel Star	20.30	−0.21	−2.04
12	Pentel BK 101-AE	19.06	0.74	−3.01
13	Pentel BK77 Superb	21.05	0.95	−2.07
14	Pentel Meteor Fine	19.97	0.13	−3.29
15	Rystor Fun STAR	20.75	0.70	−3.43
16	Pelikan Stick	21.10	0.90	−3.36
17	Zebra JIMNIE Fine	20.17	0.81	−4.32
18	Sanford Saga fine Korea	22.18	1.39	−4.08
19	Corvina “51” made in Italy	22.35	1.17	−2.39
20	Markant orange	21.15	1.00	−3.50
21	Warwick	18.37	2.19	−3.65
22	Pilot BPP-GPL-F-B	20.13	3.34	−5.54
23	BIC Diamante	20.97	1.68	−3.70
24	Zebra JIMNIE Light	21.12	1.21	−3.29
25	PILOT BPRG-10R-F-B RexGrip	18.86	3.05	−4.99
26	Pentel BK437	19.58	1.45	−3.73
27	STAEDTLER triplus ball M	21.20	1.20	−3.35
28	Penac CH6	23.16	0.63	−3.17
29	PAPER MATE Stick 2020 F	21.17	0.70	−2.62
30	PAPER MATE Click 2020 M	20.41	0.60	−2.62
31	Pentel BK77 Superb	21.41	1.03	−2.80
32	Uni Laknock Fine	20.91	0.89	−2.57
33	PILOT BPRG-10R-F-B RexGrip	20.84	1.56	−3.34
34	Zebra JIMNIE Light	21.18	0.53	−2.94
35	Pelikan Stick	22.13	1.44	−3.56
36	PAPER MATE Stick 2020 F	21.15	0.53	−3.53
37	Rystor Fine STAR	23.25	0.70	−4.23
38	Warwick	20.81	1.16	−3.54
39	BIC N-S Fine	21.51	1.39	−4.54
40	Renishaw	20.95	1.07	−3.68
41	Toma Superfine 069	23.79	0.86	−3.18
42	Patio Vigo	23.49	0.59	−3.34
43	Pentel BK-77 Superb	22.23	0.94	−2.89
44	Karin 113 BNP	22.21	1.30	−3.85
45	Lexi 5	21.82	1.31	−4.82

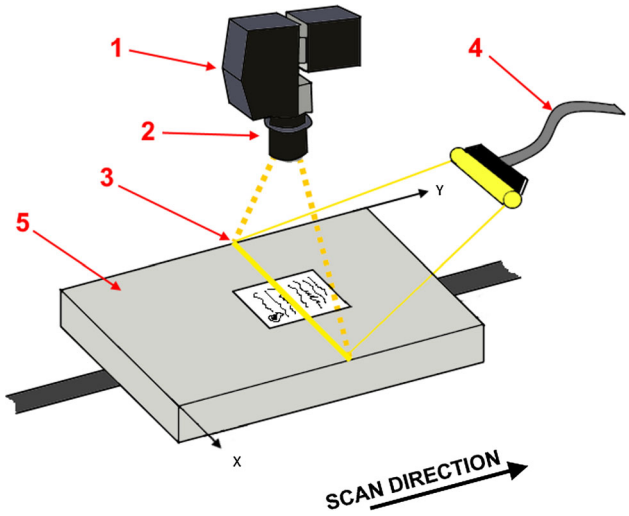


Fig. 2 Hyperspectral imaging system configuration during data acquisition: *1* hyperspectral imaging Headwall Photonics model VNIR C-series; *2* cylindrical lens; *3* light line on the sample surface; *4* fibre optic light line; *5* motorised table

intensity thresholding of the total reflectance at 550 nm. At this wavelength, the registered pens were clearly visible, while above this value, in the infrared region, ink areas become almost completely transparent. Additionally, it has to be noted that such simple threshold conditions on individual (if carefully selected) images yield results that avoid the problems caused by the “mixed” pixels at the border of an ink line. The intention of defining ROIs was to select the areas with representative and homogeneous spectral response for specific ink areas. Reflectance spectra (as mean spectra) extracted from the ROI were used to perform chemometric analysis based on hierarchical cluster analysis, principal component analysis, and spectral angle mapping, in order to establish a method of distinguishing between the inks. The procedure of registering and selecting the regions of interest was identical for the test and main set of samples.

2.4 Data processing principles

*2.4.1 CIE L*a*b colour parameters*

Colour parameters were calculated using the CIE L^*a^*b colour space system [24, 25], in which L^* is the lightness variable, and chromaticity coordinates are represented by a^* (redness/greenness) and b^* (yellowness). Colour measurements were performed using a diffuse CIE standard “D65” illuminant, at an angle of observation of 10°. Table 1 shows the mean colour data from ROIs selected in the previous step.

2.4.2 Hierarchical cluster analysis (HCA)

Hierarchical cluster analysis is one of the unsupervised methods of classification and is usually used at the beginning of research to obtain general information about the patterns present in the dataset. The approach provides a graphical representation of the existing relationships between objects. The clusters are shown in a hierarchical tree, where each case is assigned to a separate group by the distance between objects (for example, the square Euclidean distance). As the distance increases, objects are arranged in bigger clusters until a single cluster is created. Cluster analysis allows the detection of patterns in a dataset without explaining why they occur. Besides selecting a distance measure, choosing the appropriate agglomeration method is an important aspect of the study. The classification of elements for one cluster depends on the correct determination of the distance between clusters.

In this study, Ward's method and squared Euclidean distance were used. Analysis of variances led to the estimation of the distance between clusters [26]. Such an approach allows both the sum of squared deviations of any two clusters for each step of cluster analysis and the number of calculated clusters to be minimised. The hierarchical method chosen in this study produces families of clusters which themselves contain other clusters. The objective of cluster analysis was to determine similarities and dissimilarities between inks, and to categorise them. The purpose was to establish whether there is a spectral distinction between particular inks based on spectral response over the registered wavelengths.

2.4.3 Principal component analysis (PCA)

Principal component analysis is the main method of multivariate data analysis. The idea behind this approach is to allow comprehensive analysis of a dataset by reducing the dimensionality of data, and to present the patterns and data structure through new independent vectors called principal components (PC). The procedure uses orthogonal transformation to convert a system into a set of values of linearly uncorrelated variables, retaining as much variation present in the dataset as possible. The newly designated components contain all the variations present in the data, but the value decreases as the number of components increases [27]. Thus, the first component contains the most information about data variation, the second less, and so on. Principal component analysis is mainly based on the determination of the eigenvalue and eigenvectors, which are essential for further analysis and for the calculation of principal components and determining their correlation with the variables based on the factor loadings.

In this study, we provide two kinds of analysis based on the spectral information. These are the numerical approach, with the extracted spectral fingerprints of all inks, and the faster one directly in the hyperspectral image. The second approach gives, as a result, graphical data; it is time-saving and leads to similar differentiation of ink traces, making it an effective and useful method. The necessary transformations were executed using ENVI software.

The purpose of the PCA was to determine whether component spectra can result in distinguishing between the spectra of representative inks. The calculation was performed on an extracted mean spectra for every 45 ink traces, as well as on the entire hyperspectral image, to produce numerical and graphical results. Orthogonal transformation was applied to the correlation matrix in this study. This rotates results in component patterns that are more open to visual interpretation and allows loading analysis in order to estimate how the spectra of inks load to the principal component.

2.4.4 Spectral angle mapper (SAM)

Spectral angle mapper is an algorithm that permits the measurement of spectral similarity between two spectra which could be expressed on a numerical scale (from 0—no similarity to 1—identical spectra). In this approach, selected spectra are treated as vectors in n-dimensional space, in which the number of dimensions is equal to the number of recorded spectral lines. This allows the calculation of the spectral angle [28]. It is worth mentioning that the SAM method is resistant to illumination variation, which can be explained by considering two spectra as vectors with each point on the line representing the same material characteristics but a different illumination value. The calculated angle between these two vectors with the same origin is constant and independent of illumination. Smaller angles represent a closer similarity to the pattern, and the pixels outside the specified threshold of maximum angle are not classified. Similarly to the PCA approach, the data can be analysed numerically, leading to a matrix of results that combines every pair of traces, or, in a graphical approach, that shows the similarity of a selected ROI with every pixel within the image.

In this study, the first step was to establish the correct spectral angle that met the requirement of sample differentiation. This was done by determining the maximal value of the spectral angle for two identical spectra (and repeated for every ten pairs of selected identical inks). The maximum value of spectral angle was determined experimentally, based on the similarity between a pair of spectra for samples P21 and P38, the spectral responses of which in the analysed spectral range were almost identical. The

angle was determined by adjusting its value in such a way that the numerical similarity between these two samples was as high as possible, while simultaneously, the similarity between this pair of samples and other samples was lower. The maximum spectral angle determined by this method was 0.1 rads, and this value was used in further studies. In this approach, each ROI (each one corresponding to one trace of pen) was used alternately as reference spectra for SAM algorithm. That methodology was chosen because of the resistance of the calculated similarity between every two spectra to the level of illumination, and because it allows differences to be expressed as numerical values.

3 Results and discussion

Hyperspectral scanning gives the opportunity to select many image pixels within a very narrow line of trace, and to average their spectral response. The spectral characteristics of all of the pens used in this study are shown in Fig. 3a. Each spectrum represents the mean absorption of inks within previously selected ROIs. The wavelength range chosen in the study was that of the visible and near-infrared region as an attempt to cover all regions of lamp light emission with good intensity. It can be seen that the studied inks have similar spectra. The regions in which the inks vary the most are in the range of 700–750 nm, and around 650 nm. These two major differences were common for all inks. The inks spectra that are different from other and can be selected visually from the whole group were P5, P7, P15, P16, P18, P19, P28, P35, and P37. Radiation in this range is orange–red in colour and is absorbed by blue–green substances. This suggests a different amount of these pigments or dyes. These average spectra do not allow direct classification and discrimination of the samples, and nothing can be said about the differences observed in the rest of the spectra of inks because of the high level of similarity. Due to the subtle differences between the samples, simple comparison is not efficient for differentiation. The colorimetric data presented graphically in Fig. 3b do not show significant differences between samples. The distances between particular points in the graphs presented do not allow classification of the pens into separate groups. The obtained results suggest ink colours so similar that the human eye cannot see the differences. Therefore, advanced statistical tests were judged necessary for the establishment of their classification into groups.

In this study, the independent test set was prepared, and inks were examined to ascertain whether those from the same source were more similar to each other rather than to those from other sources, and to establish whether obtained results were reliable. The test included four different pen

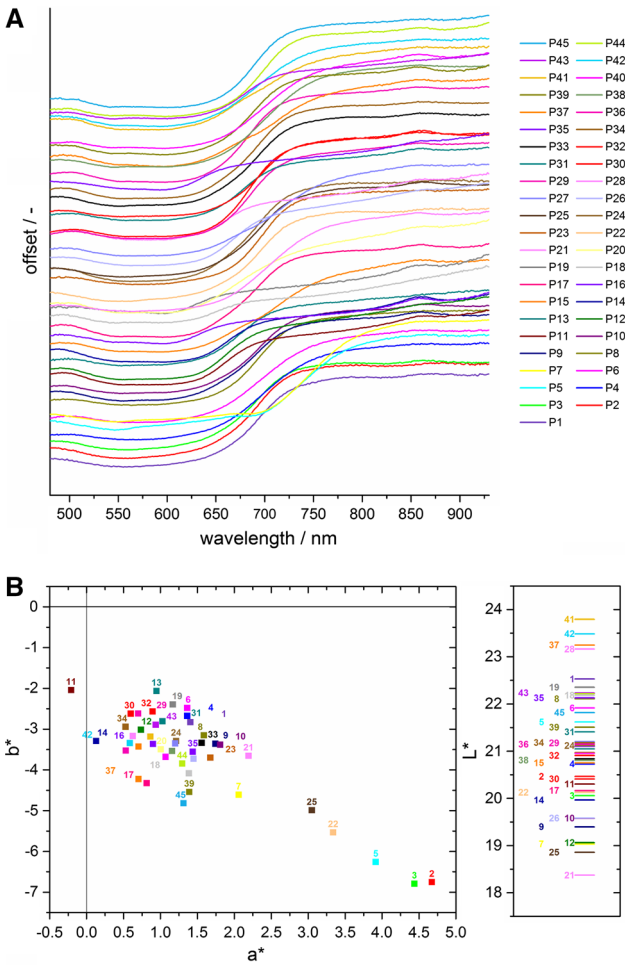
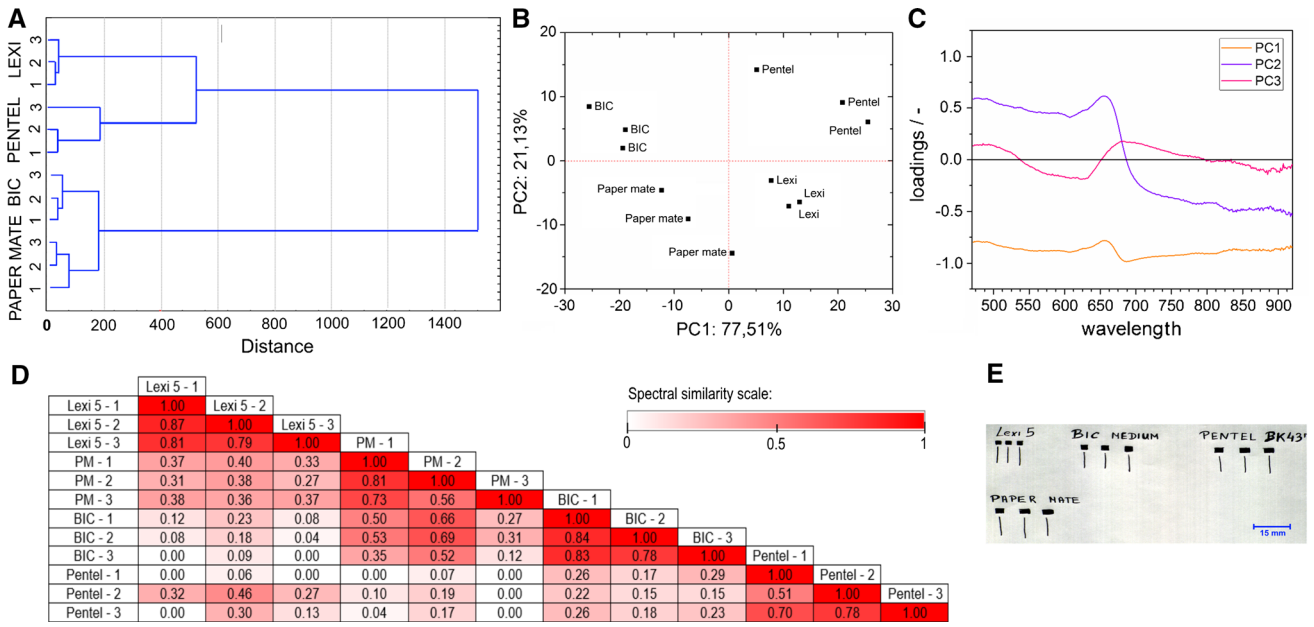


Fig. 3 **a** The average spectrum (from selected ROI) for each ink line used in this study; **b** colorimetric data calculated on hyperspectral images and extracted as a mean value from the same ROI as spectral data; each point represents the mean CIE $L^*a^*b^*$ over analysed ROI

inks (from BIC, Pentel, Lexi 5, and Paper Mate). For each pen, three lines were placed on a paper substrate. Those inks were chosen based on their spectral response, and while they have similar characteristics, it is possible to determine regions that may allow differentiation. Selected chemometric methods (HCA, PCA, and SAM) were used. The results of independent test set analysis are presented in Fig. 4. Correct classification rates of 100 % were obtained for the test set. Separation obtained by HCA agreed well with the principal component analysis results. HCA, as an unsupervised technique, not only identified four clear clusters but also showed similarities between samples. The inks vary mostly in the region around 650–700 nm, as shown in Fig. 4C.

SAM results help establish the similarities between every pair of analysed spectra. Before the calculation of final similarity/dissimilarity, the maximum spectral angle has to be selected. This was done experimentally following procedure similar to described above. A maximum spectral



	Lexi 5 - 1	Lexi 5 - 2	Lexi 5 - 3	PM - 1	PM - 2	PM - 3	BIC - 1	BIC - 2	BIC - 3	Pentel - 1	Pentel - 2	Pentel - 3
Lexi 5 - 1	1.00											
Lexi 5 - 2	0.87	1.00										
Lexi 5 - 3	0.81	0.79	1.00									
PM - 1	0.37	0.40	0.33	1.00								
PM - 2	0.31	0.38	0.27	0.81	1.00							
PM - 3	0.38	0.36	0.37	0.73	0.56	1.00						
BIC - 1	0.12	0.23	0.08	0.50	0.66	0.27	1.00					
BIC - 2	0.08	0.18	0.04	0.53	0.69	0.31	0.84	1.00				
BIC - 3	0.00	0.09	0.00	0.35	0.52	0.12	0.83	0.78	1.00			
Pentel - 1	0.00	0.06	0.00	0.00	0.07	0.00	0.26	0.17	0.29	1.00		
Pentel - 2	0.32	0.46	0.27	0.10	0.19	0.00	0.22	0.15	0.15	0.51	1.00	
Pentel - 3	0.00	0.30	0.13	0.04	0.17	0.00	0.26	0.18	0.23	0.70	0.78	1.00

Fig. 4 Results of independent test set analysis. **a** Results of HCA analysis; the four groups of samples were adequately classified; **b**, **c** results of PCA analysis; **b** PC1 versus PC2 plot—the separation of samples suggests the presence of four clusters; **c** wavelength versus loadings for the first three PCs; **d** results of SAM analysis with the spectral angle equal to 0.1 radians. The *red marked areas* indicate the samples with high similarities; **e** preview of the registered test image [an RGB image constructed from images registered at 640 (*red*), 550 (*green*), 460 (*blue*) nm]

angle of 0.1 radians was determined. The test proved the power of SAM in distinguishing the Lexi 5 inks from those of the Pentel and BIC pens, and the Paper Mate and BIC from the Pentel inks (see Fig. 4d). The HCA, PCA, and SAM methods correctly divided the spectral data into four groups consisting of three different pens. Results proved that analysis is reproducible.

The spectra included in the main set were analysed similarly. The aim of using these algorithms for these data was to differentiate between 35 various black inks. The prepared set may simulate a real case, in which the number of inks is not known. The first step in distinguishing between the black traces in the main dataset was to carry out HCA on the averaged extracted spectra from the hyperspectral image. The appropriate clustering algorithm and applied parameter settings are significant steps that must be taken before analysis, and depend on individually analysed data. The results of cluster analysis are encouraging with respect to the separability of ink reflectance. Such analysis does not focus on the selection of the highest number of groups, but on the determination of several probable clusters, the content of which is consistent with the results achieved by other methods.

The dendrogram is shown in Fig. 5. The left site indicates as many clusters as spectra, while on the right, there is only one cluster. Arbitrarily, through the clusters, the vertical red line that intersects the horizontal lines is placed at a particular Euclidean distance to establish the number of clusters present in the dataset. The obtained dendrogram (Fig. 5) can be divided into two levels of cut-off, which indicates the presence of two or three clusters. The final dendrogram shows a group of samples that are relatively close. Due to the presence of a significant gap between the first and second branch, it looks as if two very well-defined clusters are present on the dendrogram. It was noted that by

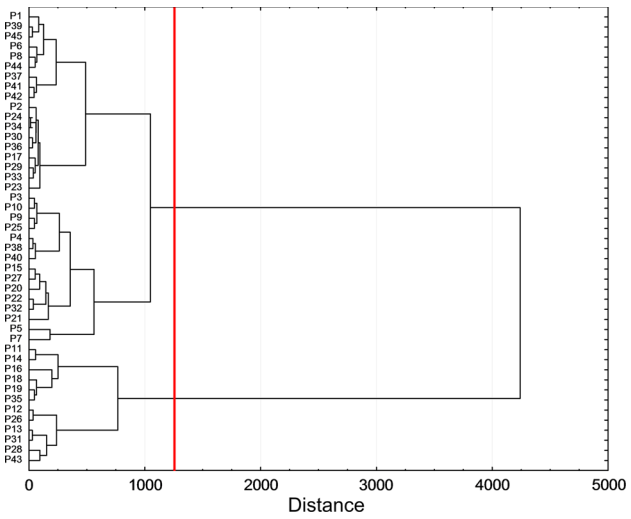


Fig. 5 Dendrogram representation of HCA results. Obtained data can be divided into two main groups at the distance of about 1250 (the *red line* represents cut-off level). The following subgroups can be correlated with other classification methods

selecting the cut-off level for a smaller bond length, some traces made by using the same ink were classified into separate clusters. The selected cut-off level indicates two clusters corresponding to a stronger separation. This was the basis for classification into two groups (Table 2). The partitioning results in two yielding clusters consisting of 33 and 12 elements. Ten pairs of traces were created using the same black pen. In Table 2, each pair is marked with the same colour. According to the results of cluster analysis, these pairs have been assigned to the same group. This separation agreed well with the principal component analysis results (see below). HCA, as an unsupervised technique, not only determined specific differences but also showed similarities between inks. Further steps of analysis were focused on detection of subgroups in order to distinguish as many pairs of traces as possible.

The PCA can be applied directly to the hyperspectral image, which is especially useful when the RGB image of the hyperspectral data does not show any visible differences between traces. Selected images of the calculated principal component are presented in Fig. 6. PC1 does not provide information about differences between analysed traces, but it can explain the difference between all writing traces and the background (most of the variance in the image is the difference between black and white as has already been mentioned). PC6 and PC8 show the distinct division of the samples into two separate groups, which corresponds to the results of cluster analysis when two traces (P5 and P7) diverged significantly from the rest.

In the next step, the PCA algorithm based on the correlation matrix was applied on the average spectra defined by ROIs. The first three principal components explain more than 96 % of all the variation in the examined dataset (PC1 47.08 %; PC2 40.84 %; PC3 8.89 %; PC4 1.35 %; PC5 1.08 %; PC6 0.34 %; PC7 0.21 %; PC8 0.08 %; and PC9 0.03 %). It is worth mentioning that the first PC represents data that differentiate the white background from the written text, and as such are irrelevant to the differentiation between 45 analysed ink traces. Representation of the original correlated variables using principal components in orthogonal space reduces the dimensionality of the system while retaining essential information about the variability in the data. As described earlier, graphical representations of PCs helped in the choice of PCs for drawing loading and scatter plots. The loadings were plotted (Fig. 7) for the PC1, PC2, PC3, PC6, and PC8. This kind of graph gives information about the correlation between loadings and variables. The interpretation of the principal components is

Table 2 Results of hierarchical cluster analysis. The two main clusters can be separated. The colour marks represent pairs of same pen media used during analysis

Cluster no.	Name of the black pen traces
1	P6, P40, P8, P39, P15, P37, P2, P32, P24, P34, P29, P36, P25, P33, P21,P38, P27, P20, P22, P5, P7, P1, P45, P44, P41, P42, P30, P17, P23, P3, P10, P9, P4,
2	P13, P31, P16, P35, P18, P19, P12, P26, P28, P43, P11, P14,

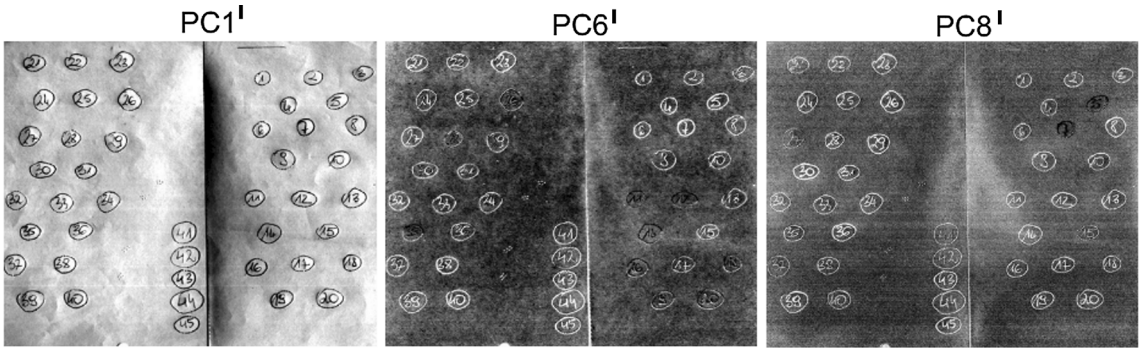


Fig. 6 Graphical representation of the principal components taken from the hyperspectral cube analysis. The first component divides the background information from the gel pen inks; the sixth and eighth principal components represent the division of the ink traces into two different groups (*white* vs. *black* traces)

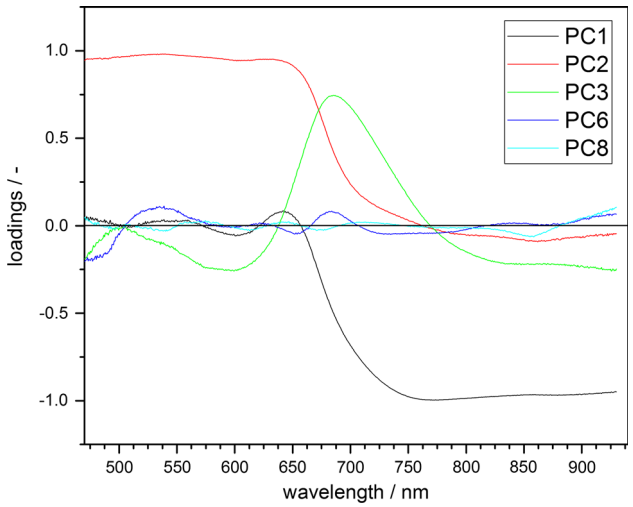


Fig. 7 Loadings plot for the first, second, third, sixth, and eighth principal components

generally based on finding which variables are most strongly correlated with each component, and how these variables are correlated with loadings. PC1 has a strong negative correlation with variables within the spectral range 700–930 nm. PC2 has a strong positive correlation between approximately 470–670 nm. PC3 has a strong positive correlation between 670 and 700 nm. In order to achieve a complete picture of the correlation of principal components with variables in that dataset, the loadings plot (Fig. 7) was compared with a scatter plot of selected principal component scores. Scatter plots of principal component scores were prepared by plotting the first, second, third, sixth, and eighth components against each other (Fig. 8). According to this analysis, the samples can be divided into two groups of traces dependent on the correlation with the principal components. Scores between the two groups are differentiated primarily by PC1. This, in combination with the loadings graph (Fig. 7), indicates that differences in the spectra occur mainly within the spectral range 700–930 nm. Additionally, the results shown in Figs. 7 and 8 suggest that pen traces P5 and P7 are negatively correlated with PC3 and are far from the other traces (both of these samples have a greater value of calculated distance), creating a separate cluster. This cluster can be distinguished from the other by the different spectral response in the range of 670–700 nm, which is associated with a strong positive correlation of variables to PC3 within the considered spectral range. This indicates that discussed pairs of traces were prepared using the same black ink or very similar in composition. These pairs could be expected to have a close correlation with the principal components, and to be within a short distance of each other on the scatter plot. This separation agreed well with the HCA results in Fig. 5, where it can be seen that samples P5 and P7 create separate branch. Thus, this pair of inks seems to be different from the rest (see Fig. 3), so data corresponding to P5 and P7 were removed and the PCA algorithm was applied to the remaining data to see if other clusters were visible. The fact that this analysis did not show additional grouping of samples suggests a strong similarity between the rest of the samples.

However, significant differences were noted in some cases, which might have been caused by a different response in the reflectance level (higher background in spectra) in spite of the fact that this effect should be compensated for by the PCA algorithm applied to averaged spectra. It is probable that an uneven distribution of media on the paper surface could also influence the results, as PCA is relatively sensitive to illumination and the albedo effect. To confirm that observation, the reflectance plot was constructed for some pairs of samples using the same gel pen inks (Fig. 9). When a pen media was applied to the white surface as a thinner layer, it had a higher reflectance value in comparison with the thicker layer. Thus, observed differences are not a result of different sample composition, but are associated with the thickness of the deposited gel layer. Unequal distribution of cellulose fibres in the paper (resulting in a non-homogeneous paper surface) could have an impact on the lower correlation between those traces.

Using the graphical representation of PCA (as presented in Fig. 6) is not always an easy way to distinguish exactly the same group comparing to numerical values (presented in Fig. 8). Graphical representation analysis is based on the recognition of the areas with the colour/shade different from the background (or other areas). This assessment is subjective. The analysis is affected by perception of very small differences in the areas colour. That sometimes may lead to wrong conclusions. To avoid that effect in this study (part with graphical representation of PCA results), only selected PCs with observed differences between inks were presented. As a result of distinguishing between samples, only those which clearly vary (are darker/brighter than others) from each other can be selected. With data processing procedure, where the graphical representation is obtained as result, as an input the whole area (that includes both inks areas and the paper background) is taken into account. This approach benefits from easy extracting of the main groups present in the dataset after computing the PCs. On the other hand, relatively long time of computing the PCs (in case with high-resolution hyperspectral image) and relatively poor information about the variation of data in the dataset can be considered as main drawbacks. Numerical results of PCA analysis should be in good agreement with graphical representation. Nevertheless, as described above, some misclassified can be presented for some samples due to wrong colour/shade recognition. Numerical representation provides more easily available

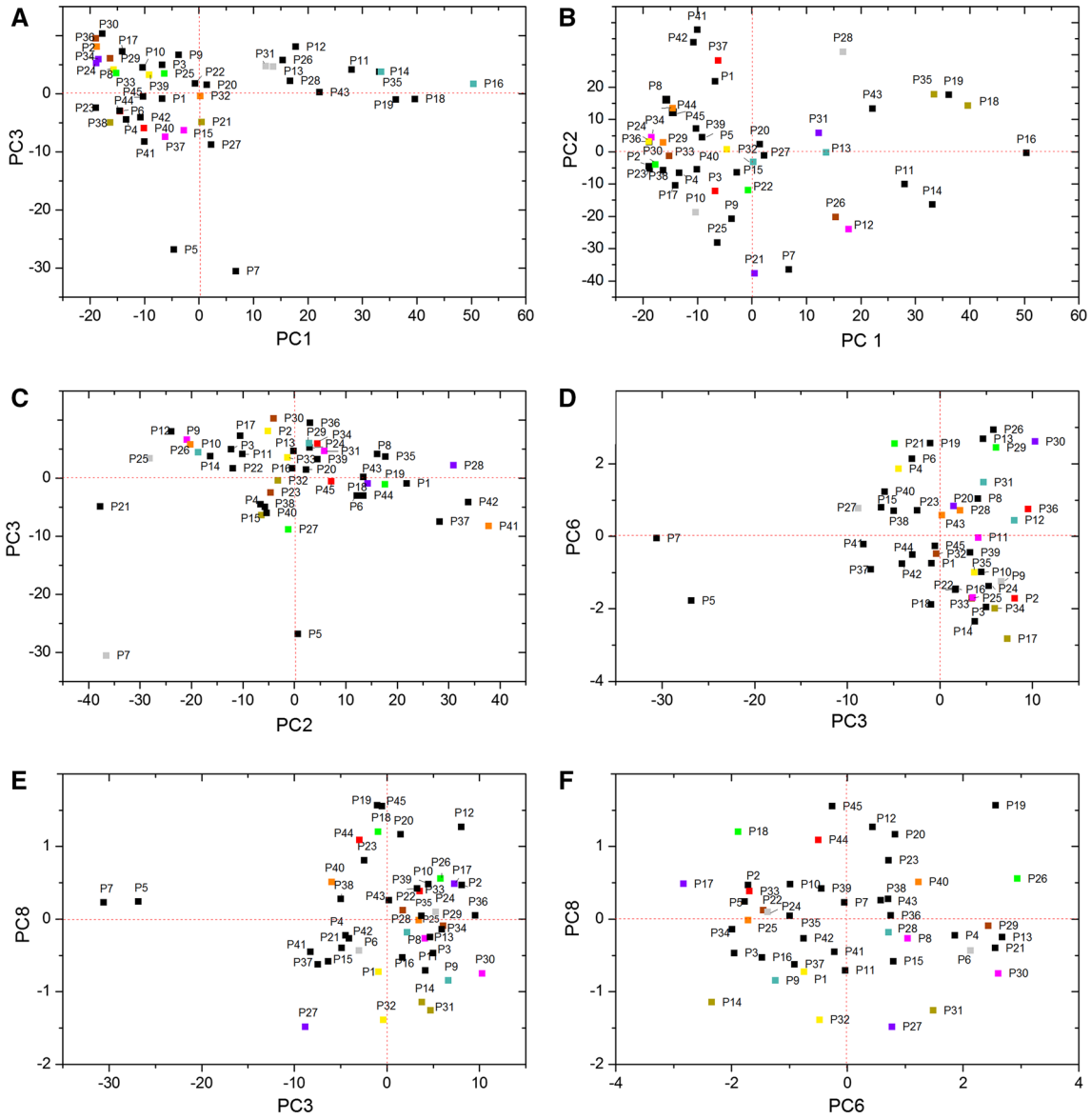


Fig. 8 Scatter plot of the selected principal components: **a** PC1 versus PC3; **b** PC1 versus PC2; **c** PC2 versus PC3; **d** PC3 versus PC6; **e** PC3 versus PC8; **f** PC6 versus PC8. *Different colours* represent pairs of identical inks used in this study

data that where each sample has its place in selected space of PCs (Fig. 8).

PCA results allowed another three groups to be determined among the clusters described by HCA and showed samples P5 and P5 to be widely separated from the main cluster. Furthermore, PCA showed the spectral range where the samples seem to vary from each other, which may be caused by different components used during ink production resulting in different levels of reflectance mostly in the range of 650–750 nm.

The third applied method was the SAM classification algorithm. It was selected due to its better performance compared to other classification algorithms [29], and because it is relatively insensitive to illumination effects. In this study, the spectra were extracted from region of interest (ROI) defined over hypercube data. Each two spectra among 45 were then tested to calculate similarity level (for each step, one was used as a reference spectrum and other as a test spectrum, resulting in 45 values of similarity for each pair of spectra on a scale of 0–1). The maximum spectral angle value was selected as described earlier. Since SAM is a supervised method, in which the operator has the ability to control the maximum spectral angle value, the results depend on the spectral angle threshold applied, the value of which should be defined during analysis. The summarised results obtained using the numerical SAM method are shown as a matrix in Fig. 10. Higher scores represent closer matches to the reference

spectrum. This approach provides a lot of information about the similarities and dissimilarities between the samples within the analysed dataset and proves that with a maximum spectral angle value of 0.1 radians, samples P5 and P7 vary from the others. The score of spectral similarity equals zero when compared to other ink spectra. A similar conclusion can be drawn from samples P35, P19, P18, and P16 which have almost identical responses to each other and differ from the rest of the samples. Moreover, HCA analysis confirms that fact, so these samples are included in one cluster.

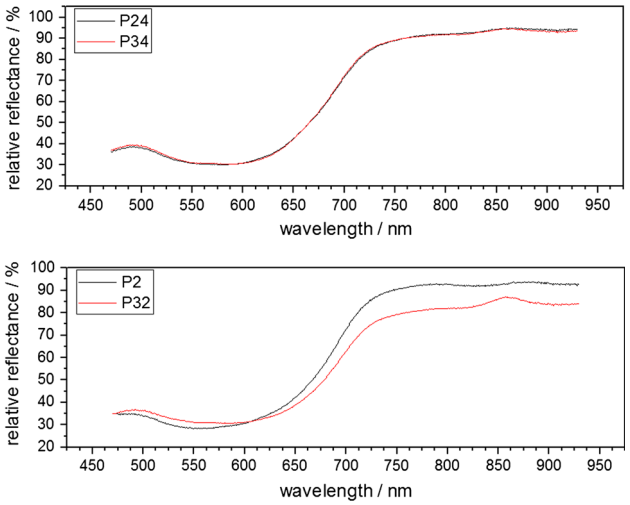


Fig. 9 Selected pairs of the same gel pen inks and their spectral characteristics in the Vis–NIR range

With SAM analysis, it was possible to determine and prove the clusters of data obtained with PCA and HCA analysis. The summarised results of differentiation are shown in Fig. 11. Optical analysis of black inks does not yield differentiation results for all inks. Indeed, such differentiation would be impossible using optical methods, due to the almost identical spectral characteristics over the analysed spectral region of 470–930 nm. This seems to be one of the limitations to this study, but it can be overcome with hyperspectral imaging working in a wider spectral range. It makes the HSI a useful alternative method of analysis for forensic purposes. The analysis of inks can

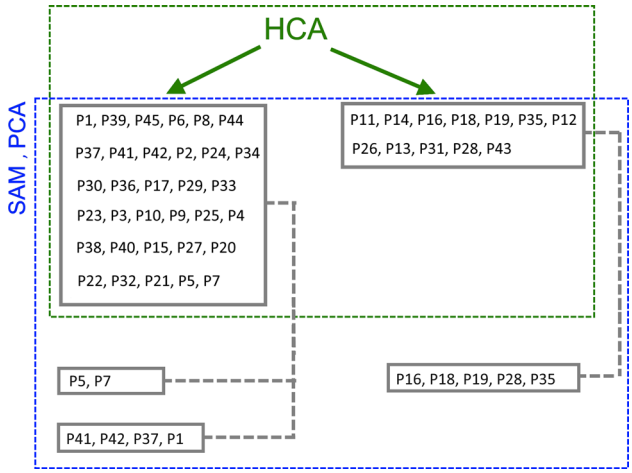


Fig. 11 Final separation of the analysed gel pen inks traces obtained by chemometric analysis

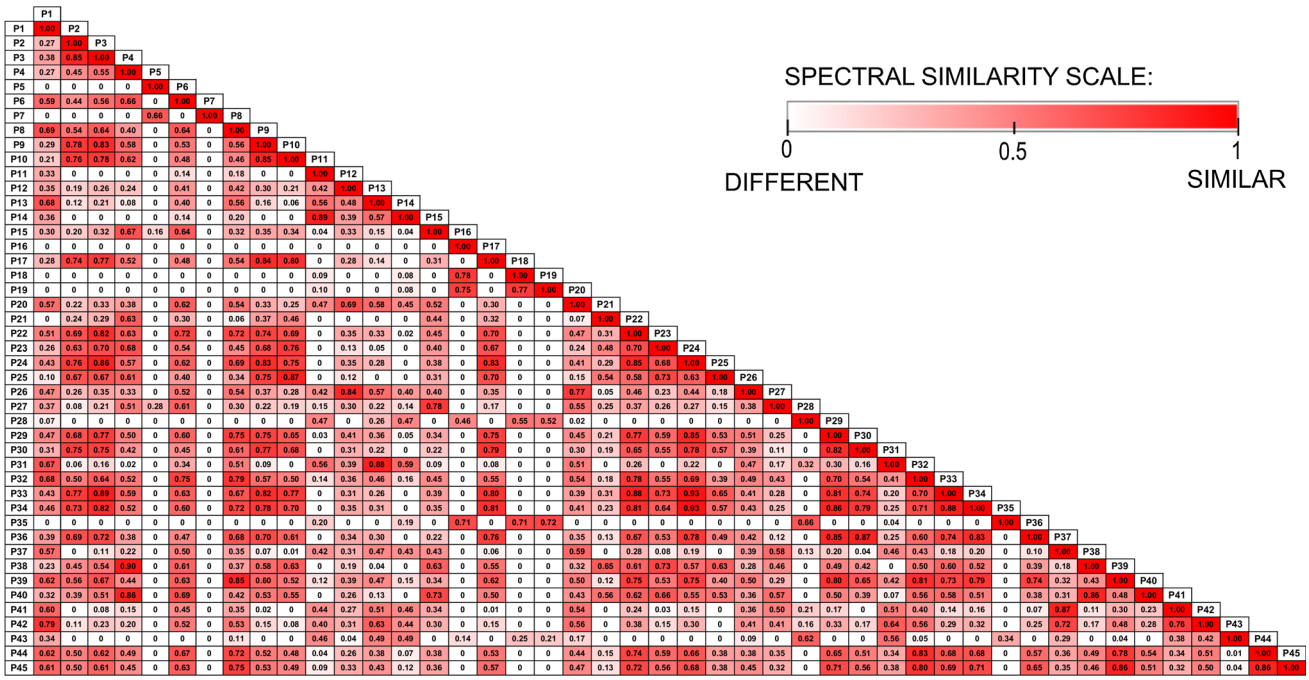


Fig. 10 Spectral angle mapper results. *Higher scores* represent closer matches of the tested spectrum to the reference spectrum

even be extended with the formation of a reference library. HSI with relatively coarse resolution and a library of inks would provide cheap and reliable classification within an entire document in a time efficient manner. Invasive methods (such as thin-layer chromatography or capillary electrophoresis) have to be repeated at several places within an ink line to give results that can be reproduced, while hyperspectral imaging allows inks to be separated from background information, and the response within selected ROIs to be averaged, all in one expeditious step.

4 Conclusions

Among the non-destructive methods of differentiation of inks, hyperspectral imaging seems to be one of the most promising tools and can be considered as an excellent alternative to invasive ones. The analysis of registered data could be based on conventional and well-known procedures, as was shown in this study. Thanks to the specific manner of measurement—registering the entire object with full spectral information within every single pixel and providing sufficient data for detailed analysis—the method is highly efficient and time-saving.

In this paper, on a study to determine the methodology for black pen analysis with HCA, PCA and hierarchical cluster analysis, principal component analysis, and spectral angle mapping were conducted using hyperspectral imagery. The results showed that combining the information extracted from the hyperspectral image with statistical data treatment can, in a non-invasive manner, provide differentiation of black inks deposited on paper. The HSI technique, in the range of 470–930 nm, allows differentiation analysis for black pen traces due to the existing differences in the chemical composition of analysed inks, mostly within spectral range 650–750 nm. Chemometric analyses allowed samples to be classified into two major separate groups and other subgroups based on HCA, PCA, and SAM analysis (see Fig. 11). The results do not allow all the inks to be distinguished, mainly due to their high similarity within the analysed spectral range, but could divide the whole set of 35 different inks into several groups. One possible way of improving these results could be UV or IR spectral range extension, as specific inks may have different reflectance responses in those regions. As shown by the colorimetric data, inks do not vary enough in colour to provide simple differentiation.

The results have been shown to be influenced not only by ink composition, but also by the degree of coverage of the tested surface. For a thin layer of the medium, spectral information of a trace may contain mixed information from the paper and the gel, which should be considered during selection of the particular regions of interest for further analysis. Nevertheless, it could be compensated for by averaging the spectral response. Gel pen inks consist of complex systems. Many additives such as dyes, pigments, resins, solvents, and emulsifiers are usually employed in ink production to provide not only colour but also other necessary features. In this study, the statistical methods were applied on mean spectra, as we know the number of different inks, but in real cases, where the spatial repartition of the different inks is not known, the selection of ROI would also be crucial. Using ROIs leads to a loss of spatial information, but it also helps to compensate for other factors such as heterogeneity, light fluctuation, and more. For unknown data, the presented methods can be applied directly on a hyperspectral image, resulting in graphical classification and allowing data quality to be quickly verified.

Another factor that affects the results of analysis is the choice of the correct spectral region. During the analyses described, such selection was limited by the imaging system configuration due to the low sensor sensitivity to UV and far-IR light. Another limitation was the light source used. The spectral power distribution of the halogen lamp does not cover spectral regions that could be helpful in distinguishing different media traces. The response from the higher wavelengths might provide results that could be more useful in distinguishing pen gels. On the other hand, working with the IR light source and very fragile documents can be dangerous for the analysed material due to heating that would occur during measurement. All these factors influence the possibility of differentiating between samples.

Acknowledgments Financial support from the Polish National Science Centre (Project No. DEC-2011/03/B/HS2/05221) is most gratefully acknowledged.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. C.S. Silva, M.F. Pimentel, R.S. Honorato, C. Pasquini, J.M. Prats-Montalbán, A. Ferrer, Analyst **139**, 5176 (2014)
2. C. Balas, V. Papadakis, N. Papadakis, A. Papadakis, E. Vazgiouraki, G. Themelis, J. Cult. Herit. **4**, 330 (2003)
3. D. Goltz, M. Attas, G. Young, E. Cloutis, M. Bedynski, J. Cult. Herit. **11**, 19 (2010)
4. C. Roux, M. Novotny, I. Evans, C. Lennard, Forensic Sci. Int. **101**, 167 (1999)
5. J. Mania, J. Bis, P. Kościelniak, Z Zagadnien Nauk Sadowych **51**, 71 (2002)

6. Z. Khan, F. Shafait, A. Mian, in *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR* (2013), p. 877
7. F. de Souza Lins Borba, R.S. Honorato, A. de Juan, Forensic Sci. Int. **249**, 73 (2015)
8. M. Claybourn, M. Ansell, Sci. Justice **40**, 261 (2000)
9. M. Hoehse, A. Paul, I. Gornushkin, U. Panne, Anal. Bioanal. Chem. **402**, 1443 (2012)
10. A. Braz, M. López-López, C. García-Ruiz, Forensic Sci. Int. **232**, 206 (2013)
11. C.D. Adam, S.L. Sherratt, V.L. Zholobenko, Forensic Sci. Int. **174**, 16 (2008)
12. A. Kher, M. Mulholland, E. Green, B. Reedy, Vib. Spectrosc. **40**, 270 (2006)
13. D.W. Sun, *Hyperspectral Imaging for Food Quality Analysis and Control* (Academic Press/Elsevier, San Diego, 2010)
14. M. Kubik, Phys. Tech. Study Art, Archaeol. Cult. Herit. **2**, 199 (2007)
15. C. Fischer, I. Kakoulli, Rev. Conserv. **7**, 3 (2006)
16. G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Forensic Sci. Int. **223**, 28 (2012)
17. A. Nakamura, H. Okuda, T. Nagaoka, N. Akiba, K. Kurosawa, K. Kuroki, F. Ichikawa, A. Torao, T. Sota, Forensic Sci. Int. **254**, 100 (2015)
18. G. Edelman, T.G. van Leeuwen, M.C.G. Aalders, Forensic Sci. Int. **223**, 72 (2012)
19. G. Reed, K. Savage, D. Edwards, N. Nic Daeid, Sci. Justice **54**, 71 (2014)
20. A. Morales, M.A. Ferrer, M. Diaz-Cabrera, C. Carmona, G.L. Thomas, in *2014 Int. Carnahan Conf. Secur. Technol.* (IEEE, 2014), pp. 1–5
21. M.J. Adams, *Chemometrics in Analytical Spectroscopy* (Royal Society of Chemistry, Cambridge, 2004)
22. K. Varmuza, P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics* (CRC Press, Boca Raton, 2009)
23. P. Buzzini, E. Suzuki, J. Raman Spectrosc. **47**, 16 (2016)
24. R.T. Marcus, AZimuth **1**, 31 (1998)
25. R.W.G. Hunt, M.R. Pointer, *Measuring Colour*, 4th edn. (Wiley, Hoboken, 2011)
26. J.H. Ward, J. Am. Stat. Assoc. **58**, 236 (1963)
27. I.T. Jolliffe, Encycl. Stat. Behav. Sci. **30**, 487 (2002)
28. F.A. Kruse, A.B. Lefkoff, J.W. Boardman, K.B. Heidebrecht, A.T. Shapiro, P.J. Barloon, A.F.H. Goetz, Remote Sens. Environ. **44**, 145 (1993)
29. J.M. Amigo, I. Martí, A. Gowen, in *Data Handl. Sci. Technol.* (2013), pp. 343–370