

Contrast in Terahertz Images of Archival Documents—Part II: Influence of Topographic Features

Tiphaine Bardon¹ · Robert K. May² · Philip F. Taday² · Matija Strlič¹

Received: 11 September 2016 / Accepted: 19 December 2016 /

Published online: 20 January 2017

© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract We investigate the potential of terahertz time-domain imaging in reflection mode to reveal archival information in documents in a non-invasive way. In particular, this study explores the parameters and signal processing tools that can be used to produce well-contrasted terahertz images of topographic features commonly found in archival documents, such as indentations left by a writing tool, as well as sieve lines. While the amplitude of the waveforms at a specific time delay can provide the most contrasted and legible images of topographic features on flat paper or parchment sheets, this parameter may not be suitable for documents that have a highly irregular surface, such as water- or fire-damaged documents. For analysis of such documents, cross-correlation of the time-domain signals can instead yield images with good contrast. Analysis of the frequency-domain representation of terahertz waveforms can also provide well-contrasted images of topographic features, with improved spatial resolution when utilising high-frequency content. Finally, we point out some of the limitations of these means of analysis for extracting information relating to topographic features of interest from documents.

Keywords Terahertz · Time domain · Imaging · Documents · Topography

1 Introduction

The ability to produce legible terahertz images of inscriptions within historical documents is dependent on the ability to discern between regions of the document with and without inscriptions. The level of contrast between such regions is dependent primarily on the

✉ Matija Strlič
m.strlic@ucl.ac.uk

¹ UCL Institute for Sustainable Heritage, The Bartlett, UCL Faculty of the Built Environment, London WC1H 0NN, UK

² TeraView Ltd, Platinum Building, St John's Innovation Park, Cambridge CB4 0DS, UK

differences in dispersive and absorptive properties between the ink used to write the inscriptions and the support underneath and around these inscriptions [1]. If a text within a document has been written using an ink that has optical properties similar to those of the surrounding support, a terahertz image would show low contrast between regions with and without inscriptions. Terahertz time-domain imaging can detect differences in the time of arrival to the detector due to variations in surface height of a measured sample. It can therefore be used to inspect, in a non-invasive way, the topography of the surface of an object, as well as the topography of layers within objects [2–5]. Well-contrasted terahertz images of a text written using ink with optical properties similar to that of the support might therefore be obtained by using information related to indentations in the support left by the writing tool that is imparted on the terahertz time-domain waveforms reflected from the surface. Similarly, terahertz time-domain imaging could potentially reveal the presence of sieve lines and watermarks intrinsic to the paper sheets within a document. Such information is of interest to curators and historians as it can help identify the date and location of manufacture of the paper sheets [6]. This paper explores the information that is imparted by topographic features commonly found in archival documents on the terahertz time-domain waveforms reflected from a document, and identifies parameters and signal processing techniques that can be used to generate well-contrasted terahertz images of such topographic features.

2 Materials and Methods

2.1 Samples

The letters ‘UCL’ were handwritten with a clean nib pen (i.e. without ink) on a contemporary print paper and on a historical parchment sheet from 1752, the latter obtained from the Institute for Sustainable Heritage Historic Reference Material Collection. The flexible nib had a central slit which would separate when the tip of the nib is pressed against the paper, with a larger handwriting pressure causing a wider opening of the slit. If dipping the nib into ink, this split of the nib allows for more ink to be deposited on the paper surface [7]. A rag paper sheet with sieve lines from the nineteenth century from the same collection was also used in this study. A sieve line is a topographic feature in the paper sheet formed during its manufacture that is left by the threads forming the sieve when covered with wet paper pulp and left to dry.

2.2 Profilometry

A portable optical contact-free profilometer (TRACEiT from Artec) capable of scanning a maximum area of 5×5 mm at a lateral and axial resolution of $2.5 \mu\text{m}$ was used to observe indentations left on the surface of the support by a clean nib pen as well as historical sieve lines. Depth profiles from indentations and sieve lines are displayed in Fig. 1a, b and c. A photograph of the entire sieve line on rag paper further analysed with terahertz time-domain imaging is also displayed in Fig. 1d: the letter ‘L’ can be distinguished.

2.3 Terahertz Time-Domain Imaging

Terahertz images were acquired in reflection mode using a commercial terahertz time-domain imaging system (TPS spectra 3000 from TeraView Ltd, Cambridge UK), equipped with a

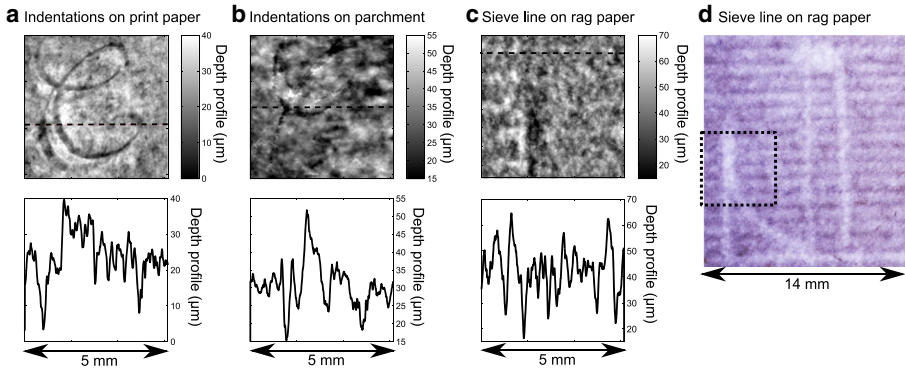


Fig. 1 **a, b** 2D and 1D depth profiles from indentations imparted on the surface of a print paper and parchment, respectively, using a clean nib pen. Images obtained using the TRACEiT profilometer. **c** 2D and 1D depth profiles of the surface sieve lines on a sheet of rag paper. 1D depth profiles correspond to depth values along the dashed line displayed on the 2D profiles. Due to the limited field of view of the profilometer (5 × 5 mm), Fig. 1a and b only contains the letter ‘C’ of the full inscriptions ‘UCL’, and Fig. 1c only displays a portion of the larger sieve line ‘L’. **d** Photograph of the sieve line obtained using the visible light transmitted through the rag paper sheet, showing the totality of the ‘L’. The dotted square indicates the area analysed with the profilometer and shown in Fig. 1c

reflection imaging module fitted to a nitrogen-purged sample compartment. This module comprises an optical system that focuses the generated terahertz pulses at an angle of incidence of 30° onto the sample surface that is sat atop a motorised scanning stage. Samples were held taut flat by placing them between two ring-shaped magnets

Spatial resolution was determined by scanning a regularly spaced two-dimensional grid (i.e. a 1951 USAF resolution test target, Product ID R3L3S1P and R3L3S1N, Thorlabs Inc.). At the limit of resolution defined by Rayleigh, two features spaced by $r = (1.22 * f * \lambda) / D$ in the scanned area are considered not resolved and the composite intensity distribution of the electromagnetic fields reflected from those features exhibits a central depression with a decrease in intensity of 19% [8]. If this definition of resolution could be translated to the amplitude of the electromagnetic field, considering that the ratio of the focal length f and aperture diameter D of the imaging system used in this study is equal to 1, it would follow that the limit of resolution of our system is approximately 180 μm at 2 THz. When scanning the grid with a lateral step size of 100 μm and producing a two-dimensional image from the amplitude at 2 THz of the Fourier-transformed reflected time-domain waveform recorded at each location of the grid, such central depression in the field amplitude was observed for features spaced by 180 μm. This figure coincides with the resolution limit set by Rayleigh. Using the same threshold for the central depression, the limit of resolution for images of the grid produced using the maximum amplitude of the reflected time-domain waveforms was reached for features spaced by 280 μm. While this threshold of 19% was arbitrarily set by Lord Rayleigh based on the capacity of the eye to discern objects through telescopes or microscopes [9], the display performances of today’s monitors, together with the advance of image processing algorithms enhancing the image contrast from inscriptions and enabling character recognition, encourage the revision of this threshold. If setting the threshold of the central depression in the electromagnetic amplitude distribution to the standard deviation of the

amplitudes collected from a blank area of the grid, a new resolution limit of 180–200 μm is reached, for images of the grid produced from the maximum amplitude of the pulse, with a lateral step size of 100 μm .

The lateral width of the indented lines left with the nib pen was measured with the profilometer and ranges between 85 and 420 μm for print paper and 85–235 μm for parchment. The lateral width of the sieve lines was measured from the photograph acquired with transmitted light and ranges between 350 and 800 μm . While the spatial resolution discussed above indicates that two indented lines spaced by less than 180 μm would not be discernible, a single indented line (farther than 180 μm from any other line) with a width smaller than the spatial resolution, or even smaller than the pixel size, may still be detected with our terahertz imaging device. Two parameters influence the detectability and contrast of any feature (i.e. single indented line in this study): the position of the feature with respect to the pixel grid (i.e. scan steps in our case) and the difference in amplitude between the electromagnetic field reflected from the feature and the electromagnetic field reflected from the surrounding background. The feature is more likely to be detected and show a good image contrast if it is fully inscribed within a single pixel, rather than overlapping two or more adjacent pixels. Plus, a large difference between the field amplitude reflected from the feature and from its surrounding background leads to an intense signal collected from the feature and a weaker signal collected from the background, resulting in a good signal-to-noise ratio and therefore to an improved detectability and contrast. When using a scanning step size (i.e. a pixel size) half the width of the indented line, there is always a pixel which is fully inscribed within the line and faithfully renders the signal from the line, so shows the maximum contrast, regardless of the position of the line with regards to the pixel grid. Using a scanning step size below half the width of the line does not improve the contrast of the inscriptions. However, selecting a smaller step size enables to render the width of the lines more accurately. To obtain an optimal contrast from the narrowest indented lines (85 μm) and ensure their detection, a step size of 40 μm should be used. This step size was not reachable with our set-up. Instead, a step size of 75 μm was used, which still enables to render an optimal contrast for lines wider than 150 μm (i.e. most of the scanned lines). For sieve line, a step size of 150 μm was sufficient to faithfully render the signal from the narrowest lines (350 μm).

A total of 4096 time points were sampled over 47.7 ps to reconstruct the waveform. Time points are therefore separated by ≈ 0.01 ps. However, for each sweep of the optical delay line, needed to record a single time-domain waveform, any jitter or drift in the waveform signal may introduce uncertainty into the measurement of the time at which the waveform maximum occurs. To determine this uncertainty, a flat gold-coated mirror was scanned with the reflection imaging module and the time at which the maximum amplitude of the waveform reflected from it occurs was measured. The standard deviation in the recorded time at which the amplitude maximum occurs is on the order of 0.05 ps, which corresponds to a distance of approximately 15 μm . It follows that only topographic features on the surface of a measured sample at a height (depth) 15 μm above (below) the nominal surface height can be reliably resolved and attributed to features associated with the document.

2.4 Signal Processing

For each measured sample, three-dimensional terahertz datasets were acquired in which the z -axis corresponds to the time domain of recorded terahertz waveforms. Corresponding two-dimensional images parallel to the plane of the scanned area (C-scans) and cross-sectional

images (B-scans) were extracted. C-scans were produced by setting the colour axis of a two-dimensional image to one of a number of quantities extracted from individual waveforms recorded at each location in the scanned two-dimensional area: maximum waveform amplitude, time delay at which the waveform maximum occurs, waveform amplitude at a specific time delay, spectral power at a single terahertz frequency, as well as correlation, covariance and cross-correlation of terahertz waveforms.

A sheet of support constitutes two interfaces with a difference of refractive indices: the air–front surface of the sheet interface and the back surface of the sheet–air interface. As a result, the incident pulse is reflected on both interfaces, with the pulse reflected on the back surface of the sheet–air interface being collected at a later time than the pulse reflected on the first air–front surface of the sheet interface. If the thickness of the sheet is comparable to the width of the incident pulse multiplied by the speed of light in nitrogen-purged air, the two reflected pulses of interest may overlap. The signal resulting from this overlap would not accurately render the position and amplitude of each of the reflected pulses of interest. To circumvent this, the sample waveform which would have been detected if the incident pulse was an infinitely narrow pulse, free from any environmental or instrumental distortions, needs to be recovered. This is also called the impulse response function (*IRF*) and it is obtained by deconvolution [10, 11]. When acquiring data in reflection mode, Fourier deconvolution consists of the inverse Fourier transform of the reflectance ratio:

$$IRF(t) = FFT^{-1} \left[\frac{FFT(s(t))}{FFT(r(t))} FFT(f(t)) \right]$$

The reference signal *r* is the signal reflected from a flat gold-coated mirror and can be assimilated to the signal from the incident pulse. Yet, due to division by the reference spectrum $R = FFT(r)$ in the calculation of the reflectance ratio, any high-frequency noise in the sample waveform *s* passing on to the sample spectrum $S = FFT(s)$ is amplified. To circumvent this, a filter *f* can be applied to the sample waveform. A double Gaussian filter was used here, as recommended by Zeitler and Shen [11]. It consists of two Gaussian filters, one high-pass (time width *HF*) and one low-pass filter (time width *LF*).

$$f(t) = e^{-\frac{t^2}{HF^2}} - e^{-\frac{t^2}{LF^2}}$$

Deconvolved signals were calculated for each scanning position (i.e. each pixel) using the built-in double-Gaussian deconvolution tool in the TVL Imaging Suite software (TeraView Ltd). The same high-frequency and low-frequency cut-offs were consistently used for all scanned samples.

One means of highlighting differences between reflected waveforms from different regions of a document that have varying topographical features, including indentations left by inscriptions, is to examine the degree of similarity between individual waveforms. Similarities between time-domain signals shifted in time were estimated by calculating the cross-correlation signal, as described by Smith [12], with the target signal being set to the waveform from the central pixel. When individual test time-domain waveforms are aligned with the target waveform, their resemblance is at a maximum, leading to a peak in the cross-correlation signal. The amplitude of this peak is maximal when the two signals have identical features, but decreases with the number of time-domain features that the two signals do not have in common. Similarly, only the cross-correlation of two identical and aligned signals would lead to a cross-correlation signal symmetrical on either side of its maximum peak, i.e. a cross-

correlation signal with both the same number of points and the same features on either side of its maximum position. A time shift between two signals sharing identical features would only break the symmetry in the number of time points on either side of the maximum peak of the cross-correlation signal. However, if two signals do not have identical features, the cross-correlation signal does not have the same features on either side of its maximum position. Therefore, both the amplitude of the maximum of the cross-correlation signal and its degree of symmetry can be used as estimates of the similarity between individual time-domain signals. Since only the symmetry of the cross-correlation signal in terms of features is of interest here, as discussed below, the cross-correlation signal was modified to set an equal number of time points n on either side of its maximum peak. The degree of symmetry S of the cross-correlation function f was calculated as follows:

$$f_+(x) = \frac{f(x) + f(-x)}{2}$$

$$f_-(x) = \frac{f(x) - f(-x)}{2}$$

$$S = \frac{\|f_+\|^2}{\|f_+\|^2 + \|f_-\|^2}$$

with the position of the maximum peak of the cross-correlation signal being set as the origin of the x -axis, where x refers to coordinates superior to the position of the maximum peak, and $-x$ to coordinates inferior to this position, and $\|f_+\|$ and $\|f_-\|$ indicate the Euclidean norm of the vectors $f_+ = [f_+(x_1), \dots, f_+(x_n)]$ and $f_- = [f_-(x_1), \dots, f_-(x_n)]$.

When analysing the frequency-domain representation of the recorded terahertz waveforms, similarities between frequency-domain spectra were estimated by calculating the covariance and correlation between the spectral power at each pixel (or scanning step) in the scanned area and the spectral power contained in the pixel at the centre of the scanned area.

Additional pulses resulting from multiple internal reflections on the sheet surfaces were detected in the waveform transmitted through each sheet of support. Such phenomenon is often called a Fabry-Perot or ‘etalon’ effect [13]. These subsequent pulses overlapped with the main pulse transmitted through the sheet and therefore changed the time delay between the main peak of the raw waveform transmitted through the sample sheet and the main peak of the reference waveform. To avoid such errors in readings of the time delay, also inducing errors in the calculation of the constant refractive index, deconvolution [10, 11] was used to separate the main transmitted pulse from the pulses originating from the etalon effect.

3 Results and Discussion

Figure 1a and b displays the surface profile of the indentations, left on paper and parchment samples, recorded by profilometry. For the paper samples, there is a $\approx 15 \mu\text{m}$ difference between the average paper surface height and the height at which inscriptions occur. The terahertz B-scan of the surface of the print paper shown in Fig. 2a clearly shows a delay of the front surface of the paper where the inscription is present due to the presence of an indentation (arrow 1) compared to an area where inscriptions are absent (arrow 2). This delay in the time of reflection from the front surface due to the presence of indentations is more clearly illustrated after deconvolution has been applied to recorded waveforms, as shown in Fig. 2c. Based on the discussion on time measurement uncertainty in the above section, in this case the

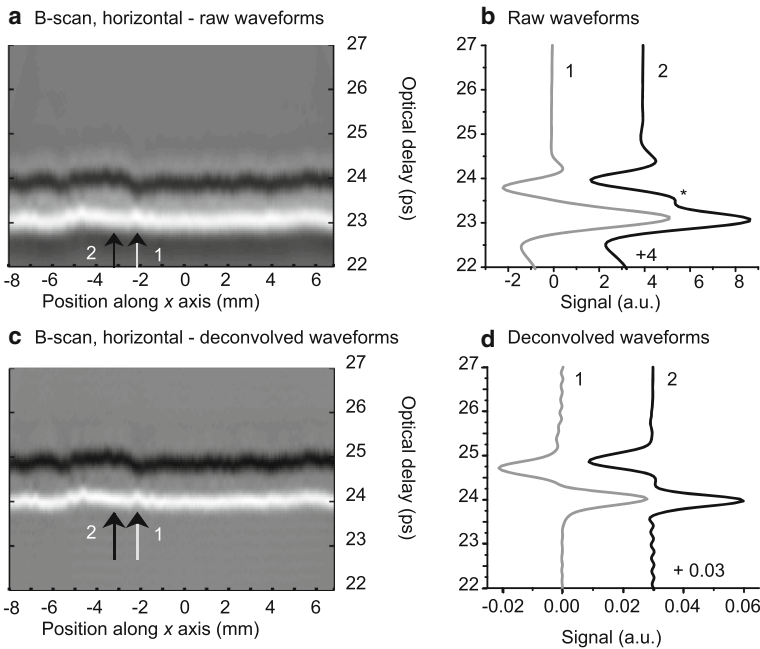


Fig. 2 **a, c** Terahertz time-domain virtual cross-section (also called B-scan) from indentations imparted on the surface of a print paper using a clean nib pen. Indentations were scanned in reflection mode with the reflection imaging module and a step size of 75 μm , along the *dotted horizontal line* shown in Fig. 3a. **(a)** and **(c)** were obtained using the raw and deconvolved terahertz waveforms, respectively. **b, d** Terahertz waveforms extracted from pixels at the positions indicated with *arrows* in **(a)** and **(c)**: the grey waveform corresponds to the presence of a surface indentation, represented by the grey arrow numbered ‘1’, and the black waveform, to the absence of surface indentation, represented by a black arrow numbered ‘2’ (this waveform is displayed with an offset of +4 in **(b)** and +0.03 in **(d)**)

depth of the indentations is large enough to reliably attribute the observed time delay of the front surface to the presence of indentations. Using this time delay, an image of the inscriptions can be constructed as shown in Fig. 3a and c: indentations left by inscriptions appear in white or light grey as their time delay is larger than for the surrounding paper, which appears in dark grey. A marginal improvement in the legibility of the inscriptions is observed when using deconvolved waveforms (Fig. 3c) rather than raw waveforms (Fig. 3a).

Figure 3d shows that no inscriptions can be observed in the terahertz image produced from the maximum amplitude of the deconvolved waveforms. This implies that the slight increase in optical path length due to the presence of an indentation (the beam travels through at least an additional $\approx 15 \mu\text{m}$ of air) does not change the amplitude of the pulse reflected at the air–paper surface interface. Moreover, the amplitude of this pulse depends on the Fresnel reflection coefficient at the air–paper interface, which is linked to the difference in refractive index between air and paper [1, 10, 14]. As a result, the fact that no noticeable change in the amplitude of the pulse reflected at the air–paper interface is detected in the presence of an indentation also implies that, under the experimental conditions used, the refractive index of paper can be considered homogeneous across the scanned area. It results that indentations do not locally modify the refractive index of paper significantly enough to be probed by our instrument. However, when using the maximum amplitude of raw waveforms to produce a terahertz image, the inscriptions are very clearly legible, as shown in Fig. 3b. Comparing the

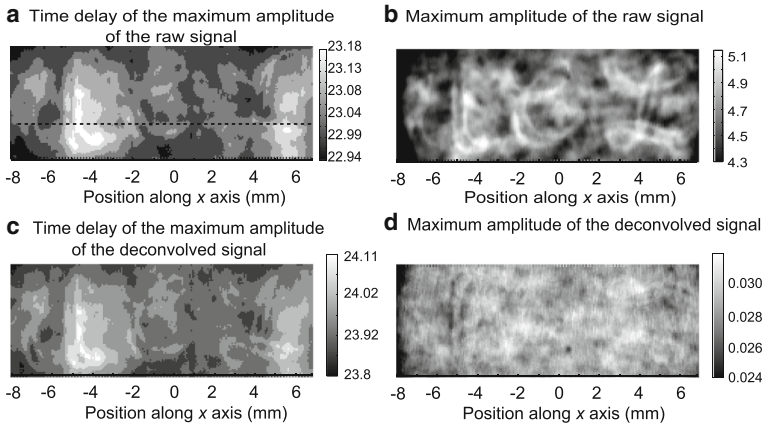
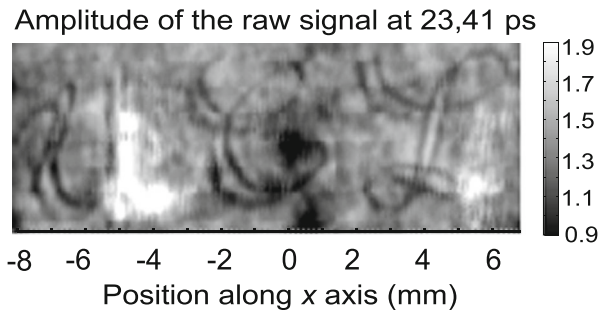


Fig. 3 Terahertz images from indentations imparted on the surface of a print paper using a clean nib pen; **a, c** the colour axis values correspond to the time position of the maximum amplitude of the raw and deconvolved waveforms, respectively; **b, d** the colour axis values correspond to the maximum amplitude of the raw and deconvolved waveforms, respectively

deconvolved waveforms (Fig. 2d) and the raw waveforms (Fig. 2b) collected from the same scanning points, it can be observed that the local thinning of paper caused by indentations brings closer the main positive pulse corresponding to the air–paper interface and the main negative pulse corresponding to the paper–air interface. In the deconvolved waveforms, this proximity does not induce any overlap between the two main pulses and therefore does not change the amplitude of the main pulse. However, in the raw waveforms, an overlap already exists between the two main pulses for the pixel with no indentation, and this overlap increases with indentation: while the overlap without indentation (*black curve*) creates a shoulder on the main pulse (indicated with a *star*), this shoulder merges with the main pulse with indentation (*grey curve*) and the amplitude of this main pulse increases. This is the phenomenon that provides the contrast observed in Fig. 3b. Interestingly, the pulse overlap is such that at the time position indicated with the *star* (23.41 ± 0.05 ps), the difference in amplitude between the waveforms with and without indentations is at a maximum. Therefore, by mapping the amplitude of the time-domain signals at this specific time position, a terahertz image with legible inscriptions can also be produced, as shown in Fig. 4. The quality of the image contrast in Figs. 3b and 4 enables to clearly evidence the line doubling caused by the splitting of the tip of the nib when handwriting with an increased pen pressure.

Analysis of raw waveforms, rather than deconvolved waveforms, therefore seems to produce clearer images of indentations at the surface of a paper sheet. In the case of a parchment sheet, the sheet itself is thicker ($170 \mu\text{m}$) than the print paper sheet ($110 \mu\text{m}$). As a result, while the indentation is still clearly observable on the B-scan in Fig. 5a (*arrow 1*), the local thinning of the sheet due to this indentation is small compared to the thickness of the sheet, and the slight change in overlap between the pulses reflected from the front and back surfaces of the parchment sheet does not induce a noticeable change in amplitude of the main pulse. It follows that the terahertz image produced using the maximum amplitude of the reflected waveform does not show any legible inscriptions (data not shown here). Secondly, the surface of the parchment sheet is more irregular (rougher and less flat) than the print paper, as observed on the profilometry image in Fig. 1b and on the B-scan in Fig. 5a. As a result, the

Fig. 4 Terahertz image from indentations imparted on the surface of a print paper using a clean nib pen, where the colour axis values correspond to the amplitude of the raw waveform at the time position giving the best image contrast



terahertz image produced using the time delay of the strongest peak (i.e. corresponding to reflection from the front surface of the parchment) does not show any legible inscriptions, as shown in Fig. 5c: the depth of indentations is similar to the variations in the surface height of the parchment sheet itself and is also similar to the depth measurement uncertainty associated with the time measurement uncertainty described in the previous section. Still, the small change in time delay induced by the indentation leads to a noticeable change in amplitude at the time position 23.04 ± 0.05 ps (corresponding to the time position of the maximum amplitude for some of the main pulses): as a result, the terahertz image produced from the

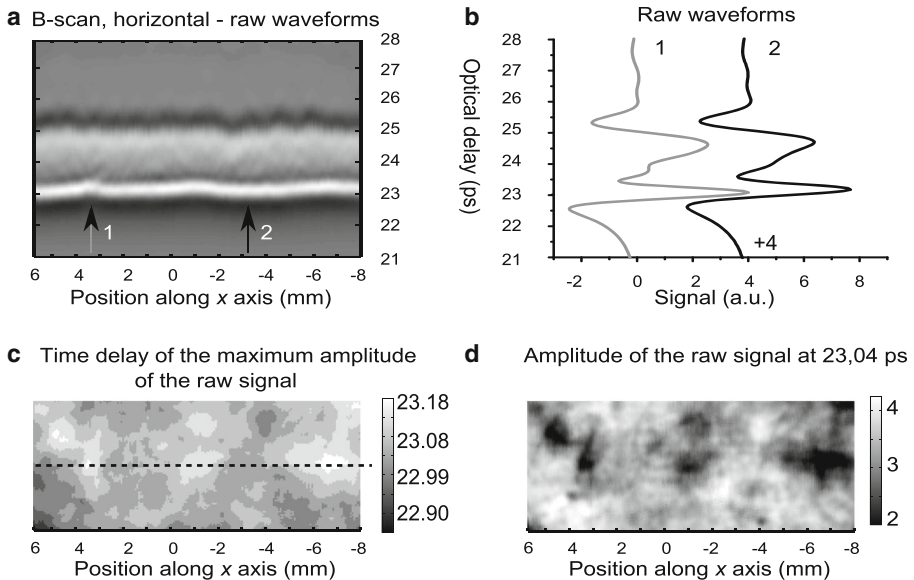


Fig. 5 **a** Terahertz time-domain virtual cross-section (also called B-scan) from indentations imparted on the surface of a parchment sheet using a clean nib pen. Indentations were scanned in reflection mode with the reflection imaging module and a step size of $75 \mu\text{m}$, along the dotted horizontal line shown in (c). **b** Raw terahertz waveforms extracted from pixels at the positions indicated with arrows in (a): the grey waveform corresponds to the presence of a surface indentation, represented by the grey arrow numbered ‘1’, and the black waveform, to the absence of surface indentation, represented by a black arrow numbered ‘2’ (this waveform is displayed with an offset of +4). **c** Terahertz image where the colour axis values correspond to the time position of the maximum amplitude of the raw waveforms. **d** Terahertz image where the colour axis values correspond to the amplitude of the raw waveform at the time position giving the best image contrast

amplitude of the signal at 23.04 ± 0.05 ps shown in Fig. 5d is the only image in which the handwritten inscriptions become at least partially legible.

In practice, documents are not perfectly flat, and in the case of water- or fire-damaged documents, they can have a very irregular and curved surface [15]. For this study, the sheets were held taut and flat by placing them between two ring-shaped magnets. However, this is not a practical solution for the study of fragile historical documents. While the terahertz images leading to the most contrasted and legible inscriptions seem to be those obtained using the amplitude of the waveforms at a specific time delay, the use of this parameter is only valid for the study of relatively flat documents. For a document with a highly irregular surface, a specific time slice would not necessarily coincide with the position of the signal reflected from the surface of the paper sheet of interest throughout the whole scanned area. As a result, the image obtained from the amplitude of the waveforms at a specific time delay is unlikely to reveal any inscriptions. Besides, it can be difficult to ascertain the time delay that will yield the highest contrast image. To overcome this issue, rather than to make use of the time difference between reflections from the surface of the sheet with or without inscriptions, the degree of dissimilarity in shape between these raw signals can be exploited, using the signal processing tool of cross-correlation. For multi-layered documents such as letters inside an envelope, the reflection corresponding to the layer of interest first needs to be identified in each waveform. In our study, this step was unnecessary since a single sheet of paper was studied. The signal to be compared to, or target signal, is the signal extracted from the central pixel, which corresponds to an area with no indentations on the paper surface in our study. All signals reflected from areas with no indentations, regardless of their time of detection, would be quasi identical to the target signal and would therefore lead to a large value of the maximum amplitude of the cross-correlation signal. On the contrary, all signals reflected from areas with indentations would not exhibit the shoulder contained in the target signal (as indicated by the *star* in Fig. 2b) and would therefore lead to a lower value of the maximum amplitude of the cross-correlation signal. Figure 6a shows the image produced from the maximum amplitude of the cross-correlation signal, for indentations left on a print paper: the inscriptions are mostly legible. The degree of symmetry of the cross-correlation signal on either side of its maximum peak can be used to produce a more legible image of the inscriptions, as shown in Fig. 6b: signals from areas with no indentations would be identical to the target signal and would lead to a cross-correlation signal with a larger degree of symmetry than the cross-correlation signal for waveforms reflected from areas without indentations. Cross-correlation is, however, only useful if there is a noticeable difference in shape between signals. Using the deconvolved waveforms reflected from the print paper, the changes in waveforms from one pixel to another is equivalent to the differences in the waveforms between areas with and without inscriptions. As a result, cross-correlation does not enable retrieval of legible inscriptions, as shown in Fig. 6c and d. The same phenomenon is observed for parchment, using either raw or deconvolved waveforms (data not shown here).

For the sample containing sieve lines on a historical rag paper sheet ≈ 100 μm thick, the same phenomena as those observed for the indentations on the sheet of print paper are observed. The thinning of paper along a sieve line, indicated by a *grey arrow* in Fig. 7a, causes a decrease in the separation between reflections from the front and back of the sheet. In the raw waveforms, this causes the shoulder band indicated by a *star* in Fig. 7b (resulting from the overlap between the pulses reflected on the front and back of the sheet) to merge with the first reflection peak, thus increasing the amplitude of the first reflection peak. It follows that individual sieve lines can be identified in Fig. 7c by using the maximum amplitude of the raw

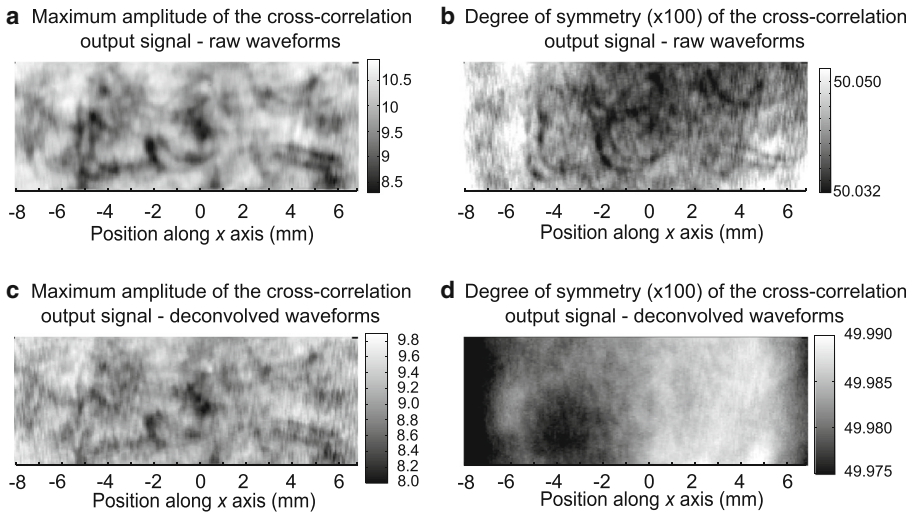
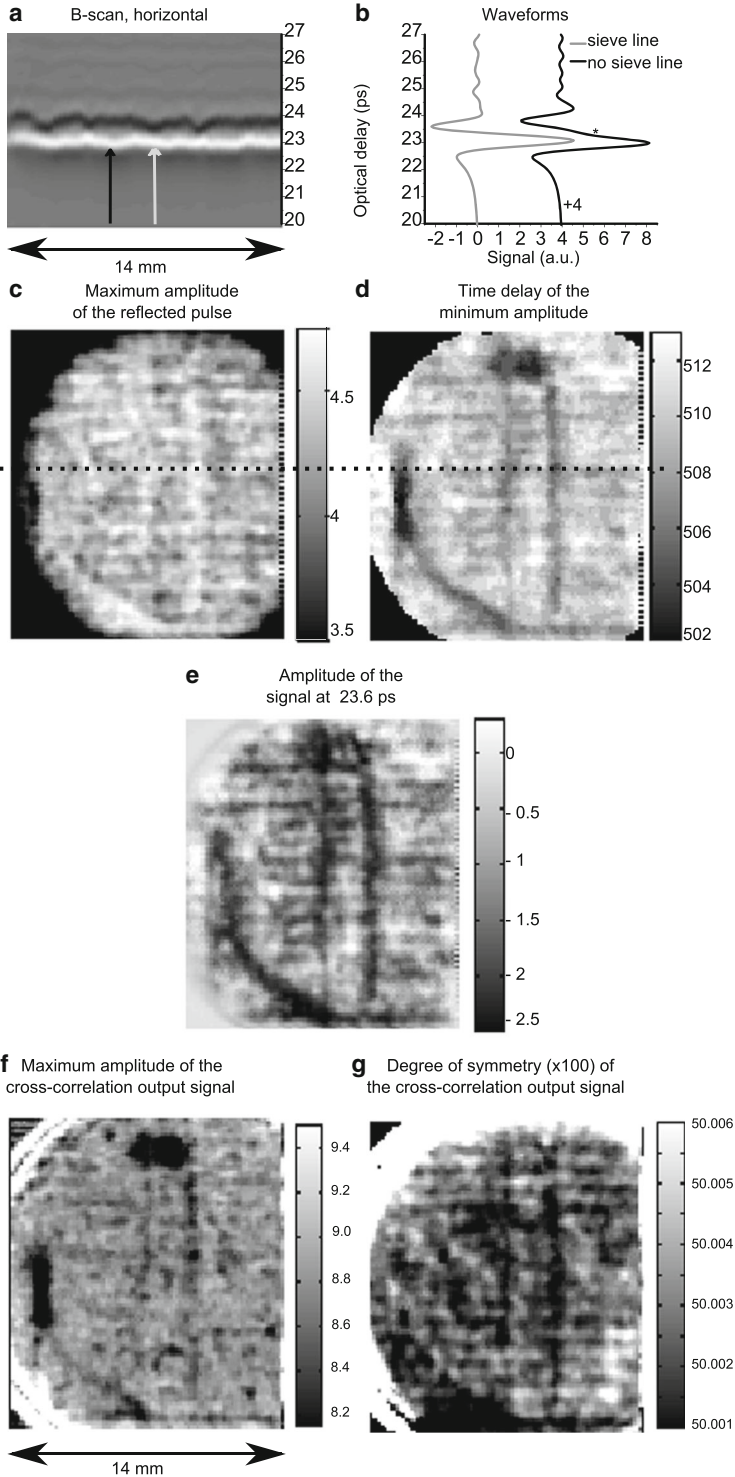


Fig. 6 Terahertz images from indentations imparted on the surface of a print paper using a clean nib pen; **a, c** the colour axis values correspond to the maximum amplitude of the cross-correlation signal, using raw and deconvolved waveforms, respectively; **b, d** the colour axis values correspond to the degree of symmetry of the cross-correlation signal, using raw and deconvolved waveforms, respectively

waveforms as the values of the colour axis in a two-dimensional C-scan. After deconvolution has been applied, the front and back reflections are more clearly resolved in time, thus lessening the influence of the presence of the second reflection peak on the amplitude of the first but therefore not encoding the presence of sieve lines on the amplitude of the first reflection peak (data not shown). For images produced using time delay or an amplitude at a specific time delay, sieve lines are also observed (see Fig. 7d and e), and the use of the raw or deconvolved waveforms matters little as long as paper thinning induced by a sieve line is more pronounced than surface roughness or curvature of the paper sheet. By making use of the absence or presence of a shoulder band in the signal, depending if this signal was reflected from an area with or without sieve line, cross-correlation efficiently produces images with clearly observable sieve lines, as shown in Fig. 7f and g.

Another alternative to overcome limitations of analysis of time-domain waveforms reflected from documents that have highly curved and/or irregular surfaces is to analyse the frequency domain content of recorded signals. Figure 8a and c displays the spectral power for each pixel (x-axis) along the dashed line indicated in Figs. 3a and 7c. Figure 8a and c is therefore the frequency-domain equivalent of the time-domain B-scans displayed in Figs. 2a and 7a. Figure 8b and d shows the spectral power signal at pixels with and without indentations or sieve lines. It can be observed that the frequency domain signal from a paper sheet changes in the presence of indentations or sieve lines. This change is induced by the change in shape of the time-domain waveforms caused by an enhanced degree of overlap between the pulses reflected from the sheet interfaces where the sheet is thinner. It is not a priori related to any change in the vibrational properties of the paper sheet caused by the presence of indentations. As a result, there are individual frequencies (or spectral ranges) where the difference in spectral power from areas with and without indentations (or sieve lines) is at a maximum. Two such frequencies are indicated by the horizontal dashed lines in Fig. 8a and c. Figure 9a and c corresponds to terahertz images produced by plotting the reflected power at those frequencies. The inscription ‘UCL’ can no longer be clearly distinguished in Fig. 9a,



◀ **Fig. 7** Terahertz signals and images acquired from the scan in reflection mode of a rag paper sheet from the nineteenth century, with a step size of 150 μm . **a** B-scan along the dotted horizontal line shown in **(c)** and **(d)**; **b** raw waveforms extracted from the pixels at the positions indicated with arrows in **(a)**, the grey curve corresponds to the presence of a sieve line, represented by the *grey arrow* in **(a)**, and the black waveform, to the absence of sieve line, represented by a *black arrow* in **(a)** (this waveform is displayed with an offset of +4); **c**, **e** terahertz images from sieve lines, where the colour axis values correspond to the maximum amplitude of the raw waveforms, the time position of the minimum amplitude of the raw waveforms and the amplitude of the raw waveforms at the time position leading to the best image contrast, respectively; **f**, **g** terahertz images where the colour axis values correspond to the maximum amplitude and the degree of symmetry of the cross-correlation signal, respectively

compared to Fig. 3b. However, the presence of sieve lines is indicated in the frequency domain signal from rag paper more significantly than indentations on print paper: the image of sieve lines produced from the difference in power at 1.75 THz shows greater contrast than the images presented in Fig. 7e, with an additional advantage that the resulting image also has improved

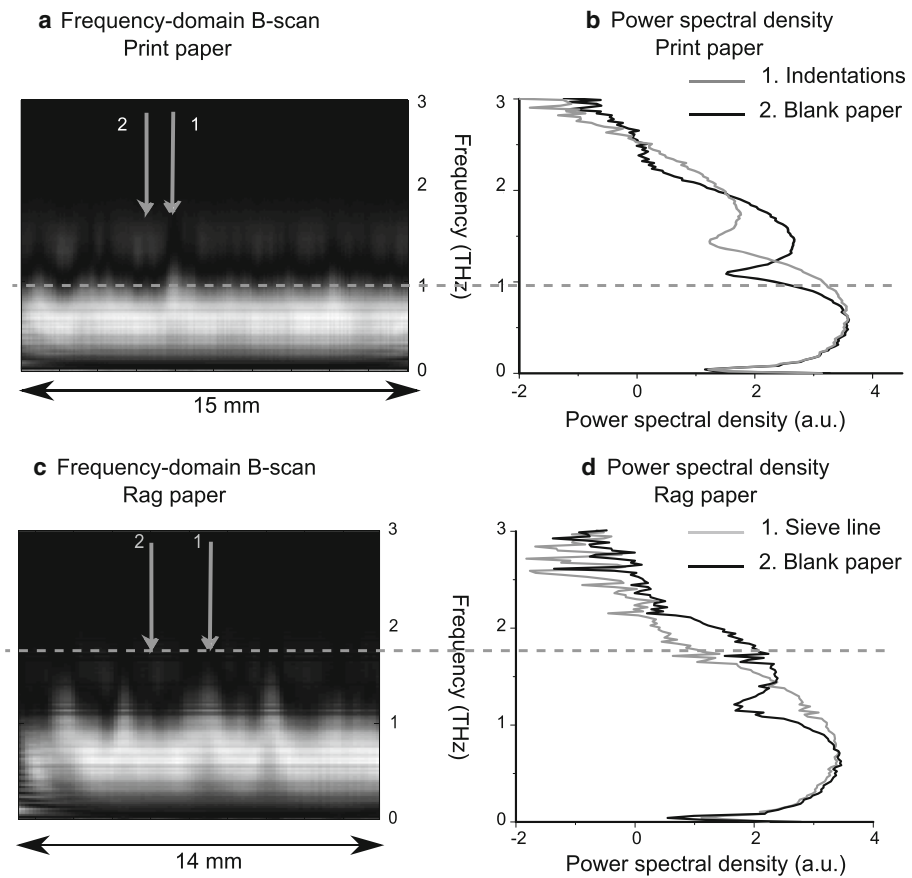


Fig. 8 **a**, **c** Terahertz frequency-domain B-scans from indentations imparted on the surface of a print paper and from sieve line on the surface of a rag paper, respectively. These B-scans were produced from data collected along the same horizontal line as indicated in Figs. 3a and 7c. **b**, **d** Spectral power at the pixels indicated by the two *vertical arrows* in **(a)** and **(c)**. The grey curve corresponds to the presence of indentations or sieve lines at the surface of the sheet (*arrow ‘1’*); the black curve, to the absence of indentations or sieve line (*arrow ‘2’*)

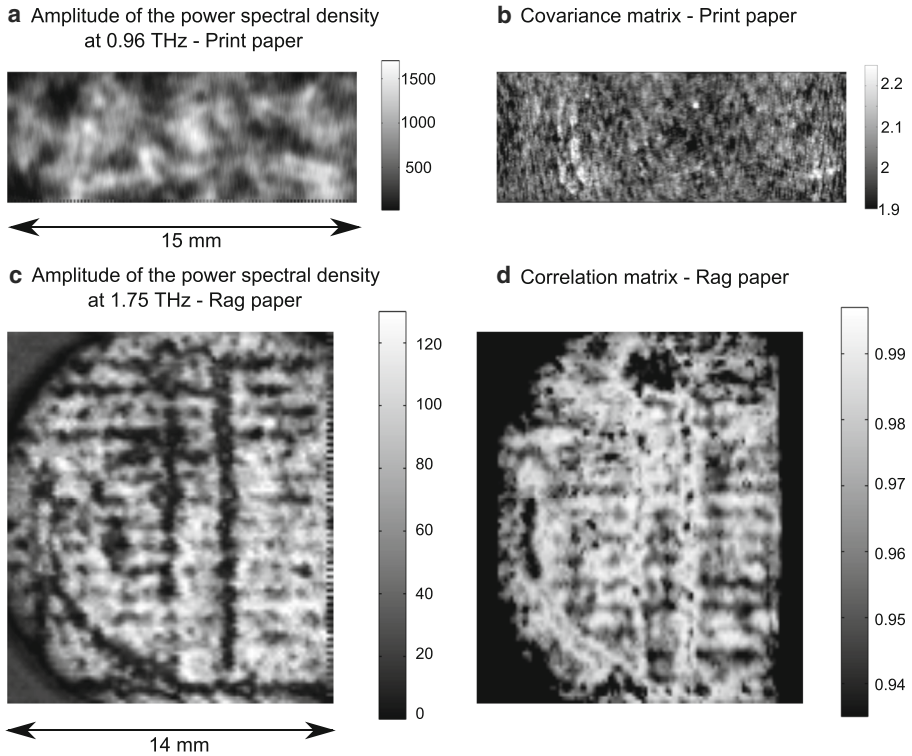


Fig. 9 **a, b** Terahertz images from indentations imparted on the surface of a print paper using a clean nib pen, where the colour axis values correspond to the spectral power at 0.96 THz and to the covariance of the spectral power, respectively. **c, d** Terahertz images from sieve lines on the surface of a rag paper sheet, where the colour axis values correspond to the spectral power at 1.75 THz and to the correlation of the spectral power, respectively

spatial resolution, due to the use of only a high-frequency component of the signal to generate the image. To avoid testing several frequency slices to find the specific frequency leading to a maximum image contrast, the shape of frequency-domain signals can be compared using correlation or covariance. The use of cross-correlation is not required here since there is no shift between the frequency-domain power spectra. Sieve lines are clearly observed in the image shown in Fig. 9d, which was produced using the correlation between the frequency-domain spectra from each pixel and the spectrum from the target pixel (which, in this case, led to better image contrast than when using the covariance—data not shown here). The image of the ‘UCL’ inscriptions in Fig. 9b was produced using covariance (improved image contrast than when using the correlation—data not shown here): as in Fig. 9a, the inscriptions are not well contrasted. Nonetheless, double-lined indentations caused by the splitting of the tip of the nib when writing with an increased pen pressure can still be observed for the letters ‘U’ and ‘C’ in Fig. 9b.

4 Conclusion

Topographic features not only influence the position of the terahertz time-domain signal reflection from the air-surface interface of a document but also its shape and peak amplitude when analysing raw waveforms. Such changes due to overlap between front and back reflections in reflected time-

domain waveforms can be used to produce well-contrasted terahertz images of indentations, sieve lines and watermarks (now shown here) not only on the surface of documents, but also inside sealed documents. The degree of overlap between the reflections from the front and back surfaces of a sheet depends on the width of the incident terahertz pulse since the detection of narrow features in the time-domain waveform, such as shoulders, depends on the bandwidth of the terahertz time-domain imaging system. However, the means by which one can obtain well-contrasted terahertz images of topographic features, such as through the use of cross-correlation or analysis of frequency-domain spectra, is system independent. Furthermore, while the use of deconvolution is instructive in interpreting the content of reflected time-domain waveforms including, for example, the identification of the surface and other interfaces, it has been shown here that information crucial to the production of legible images from topographic features may be lost after deconvolution has been applied. Finally, frequency-domain analysis provides yet another means of producing images of topographic features and can yield improved spatial resolution in the case where information on such features is encoded in the higher frequency content of the recorded signal. That being said, changes in time-domain signals due to a thinning of the sheet do not systematically lead to clearly noticeable changes in the Fourier-transformed frequency domain signals. Therefore, images produced using frequency-domain signals would not necessarily provide a significantly improved legibility of the topographic features of interest. Still, since any change in Fourier-transformed frequency-domain signals is the result of the thinning of the sheet and the subsequent overlap of the reflected time-domain pulses, and is not a priori linked to any change in the vibrational behaviour of the support, only terahertz time-domain imaging could reveal these indentations and sieve lines in the frequency domain. Continuous waves imaging in reflection mode is unlikely to lead to any contrasted image of these inscriptions.

Acknowledgements The authors would like to thank the partners TeraView, Ltd and the Nationaal Archief. Professor Sir Michael Pepper, UCL Department of Electronic and Electrical Engineering, and Professor Adam Gibson, UCL Department of Medical Physics and Bioengineering, are gratefully acknowledged for their advice. UK Engineering and Physical Research Council Doctoral Training Centre in Virtual Environments, Imaging and Visualisation is acknowledged for the financial support. This work was additionally supported by the Engineering and Physical Sciences Research Council (grant number EP/G037159/1). The authors are also grateful to Dr. Lik-Ren Tai, UCL Institute for Sustainable Heritage, for his help in acquiring depth profiles from sieve lines.

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. T. Bardon, R. K. May, B. Jackson, G. Beentjes, G. de Bruin, P. F. Taday, and M. Strlič, *Contrast in terahertz images of archives - Part I: influence of the optical parameters from the ink and support*, J. Infrared, Millimeter, Terahertz Waves [accepted], (n.d.) doi:[10.1007/s10762-016-0351-0](https://doi.org/10.1007/s10762-016-0351-0).
2. Y. C. Shen and P. F. Taday, *Development and application of terahertz pulsed imaging for nondestructive inspection of pharmaceutical tablet*, IEEE J. Sel. Top. Quantum Electron. 14, 407 (2008).
3. Q. Zhang, L. F. Gladden, P. Avalle, J. Axel Zeitler, and M. D. Mantle, *Terahertz pulsed imaging and magnetic resonance imaging as tools to probe formulation stability*, Pharmaceutics 5, 591 (2013).
4. W. L. Chan, J. Deibel, and D. M. Mittleman, *Imaging with terahertz radiation*, Reports Prog. Phys. 70, 1325 (2007).

5. J. Takayanagi, H. Jinno, S. Ichino, K. Suizu, M. Yamashita, T. Ouchi, S. Kasai, H. Ohtake, H. Uchida, N. Nishizawa, and K. Kawase, *High-resolution time-of-flight terahertz tomography using a femtosecond fiber laser*, Opt. Express 17, 7549 (2009).
6. C. R. Johnson, W. A. Sethares, M. H. Ellis, S. Haqqi, *Hunting for Paper Moldmates Among Rembrandt's Prints: Chain-line pattern matching*, IEEE Signal Process. Mag. 32, 4 (2015).
7. E. Winters, *Italic and Copperplate Calligraphy: The Basics and Beyond* (Courier Corporation, 2013).
8. E. Hecht, *Optics*, 4th edn. (Addison-Wesley, 2002)
9. Baron Rayleigh, *Investigations in optics, with special reference to the spectroscope* in Sci. Pap. Vol. I (Cambridge University Press, 1899), p. 420.
10. G. C. Walker, J. W. Bowen, J. Labaune, J.-B. Jackson, S. Hadjiloucas, J. Roberts, G. Mourou, and M. Menu, *Terahertz deconvolution*, Opt. Express 20, 27230 (2012).
11. J. A. Zeitler and Y.-C. Shen, *Industrial Applications of Terahertz Imaging*, in *Terahertz Spectrosc. Imaging*, edited by K.-E. Peiponen, J. A. Zeitler, and M. Kuwata-Gonokami (Springer, 2013), pp. 451–489.
12. S. Smith, *Digital Signal Processing: A Practical Guide for Engineers and Scientists* (Newnes, 2013).
13. G. P. Kniffin, S. Schecklman, J. Chen, S. C. Henry, L. M. Zurk, and B. Pejcinovic, *Measurement and Modeling of Terahertz Spectral Signatures from Layered Material*, SPIE Proceedings Vol. 7687, Active and Passive Signatures, edited by G. C. Gilbreath; C. T. Hawley, (2010).
14. K.-E. Peiponen, J. A. Zeitler, and M. Kuwata-Gonokami, *Terahertz Spectroscopy and Imaging*, (Springer-Verlag Berlin Heidelberg, 2013).
15. K. Pal, M. Terras, and T. Weyrich, *Interactive Exploration and Flattening of Deformed Historical Documents*, Comput. Graph. Forum 32, 327 (2013).