Application of terahertz spectroscopy for character recognition in a medieval manuscript

Kaori Fukunaga1a), Yuichi Ogawa2, Shin’ichiro Hayashi2,3 and Iwao Hosako1
1National Institute of Information and Communications Technology, Koganei, Tokyo 184–8795, Japan
2 Tohoku University, Sendai, Miyagi 981–8555, Japan
3 Riken, Sendai, Miyagi 980–0845, Japan
a) kaori@nict.go.jp

Abstract: Terahertz (THz) technology is a focus of attention in research on applied optics. We have applied THz spectroscopy and THz-imaging method to text recognition of a medieval manuscript made from sheepskin. Based on the database which contains more than 200 spectra of art materials, the red ink on the manuscript was estimated as Cinnabar. The red ink text and stains on the text were successfully distinguished by the THz-Time Domain Spectroscopy (TDS) imaging with a component spatial pattern analysis. THz spectroscopy can be used as a non-invasive analysis method for conservation science of cultural properties.

Keywords: terahertz spectroscopy, character recognition, non-invasive analysis

Classification: Science and engineering for electronics

References

1 Introduction

Terahertz (THz) technology is a focus of attention in research on applied optics [1]. Its applications include illegal drug detection through envelopes, observation of vibration modes in biological macromolecules [2], and quality control of pharmaceutical tablets. Conventional THz spectroscopy uses the Fourier transform infrared spectrophotometer (FTIR) technique. Although information in the time domain, such as phase, cannot be obtained, THz-FTIR is useful for collecting ‘fingerprint’ spectra in databases for various applications. THz time-domain spectroscopy (THz-TDS), on the other hand, measures the electric field of the transmitted or reflected waves; it provides the phase and amplitude changes of the radiation so that the refractive index and the absorption coefficient of the sample can be obtained. By using the information in the time domain waveform, THz-TDS can be used to investigate multi-layer specimens and composites that require quantitative observations in the frequency range approximately from 0.5 to 3 THz. There are also THz imaging systems that use THz-TDS. We have been using THz spectroscopy for art conservation science and have collected more than 200 spectra of pigments and binders. The availability of fingerprint spectra of various pigments suggests that THz spectroscopy can be used for non-invasive analysis of art objects [3]. This paper shows that THz spectroscopy can distinguish stains and pigments in a medieval manuscript and that a THz-TDS system together with component analysis [4, 5, 6] can be used to reconstruct a text.

2 Medieval manuscript on parchment and the specimen in this work

Parchment is a thin sheet made from calfskin, sheepskin or goatskin used for manuscripts from ancient era. It is stretched, scraped, and dried under tension. Common materials for inks were iron-gall-ink (containing FeSO₄) for black, minium (containing Pb₃O₄ or 2PbO-PbO₂), cinnabar (containing HgS), and cochineal (containing carmine acid, C₂₂H₂₀O₁₃) for red, lapis lazuli ((Na,Ca)₈(AlSiO₄)₆(S,SO₄,Cl)₁₋₂) and azurite (containing Cu₃(CO₃)₂(OH)₂) for blue, and other natural materials. Arabic gum and egg were common binders. Text recognition is an important aspect of the research on historic literature because centuries-old manuscripts are often very deteriorated. X-rays can detect metal elements, such as Fe, and therefore, conservators can estimate pigments from knowledge in the literature. On the other hand,
THz spectra are considered to be more material oriented, and hence, they should be able to indicate the compound itself. The specimen used in this work (Figure 1) was a binding fragment from the 13th century, supplied by Bernard Quaritch Ltd, UK. The observation area is indicated as a white square containing red text with some brown stains on it.

![Specimen: parchment from the 13th century.](image)

**Fig. 1.** Specimen: parchment from the 13th century.

### 3 Determination of pigments

The terahertz spectrum of the red ink was observed by the FTIR transmission system. The THz-FTIR system used in this work utilises a silicon beam splitter in the Michelson interferometer which provides wide frequency range as approximately from 0.5 to 15 THz, with a pyro-detector operating at room temperature [7]. The spectrum was compared with entries in a database containing more than 200 spectra of pigments and binders: THz Spectral Database (http://www.thz-spectra.com/). As shown in Figure 2 (a), the spectrum of the red ink in the specimen had absorption peaks around 1.1 and 3.8 THz. This spectral feature reveals that the red pigment is cinnabar (Figure 2 (b)), not minium (Figure 2 (c)).

The blue colour was difficult to be analysed by FTIR transmission system because it appeared only as a thin line in the observation area. Highly possible blue pigments are azurite or lapis lazuli, and as shown in Figure 3, neither has any particular absorption in the frequency range from 0.5 to 2 THz that is used by THz-TDS imaging system. Thus we focused red ink and stain in the observation area.

### 4 Component spatial pattern analysis for text extraction

Transmission multi-spectral terahertz images of the specimen were taken with a THz-TDS system developed by Riken. This system achieves surface plasmon terahertz imaging with metallic plasmonic crystal. [8] The spectra were measured from 0.025 to 2 THz with 25 GHz resolution, resulting in 79 images. The spectra from 0.5 to 2 THz were found sufficient to distinguish substances to perform text recognition. Figure 4 shows absorbance spectra of red ink, stain, and the base parchment, and seven frequencies indicated in the figure were used for the component spatial pattern analysis described in the following paragraph. The specimen was raster scanned over a $18 \times 28 \text{mm}^2$ area, as marked in Figure 1, which corresponds to $73 \times 113$ pixels. By calculating
time domain signals by each pixels, we obtained more than 60 multi-spectral images.

The seven images shown in Figure 5 contain three dimensional data, i.e. the position in the plane and the absorbance at the position. To extract the
distributions of ink and stain in the specimen, we compress the three-dimensional data to matrix $I_{NL}$, which is an $N \times L$ matrix, where $N$ is the number of frequency components ($N=7$ in this case) and $L$ is the number of pixels ($L=8249$). When the specimen consists of $M$ substances, featured spectral matrix of $S_{NM}$ can be obtained. The spatial distribution of each substance $P_{ML}$ is then calculated using the component spatial pattern analysis [8-10] as, $[P_{ML}]=[S_{NM}]^{\dagger}[S_{NM}]^{-1}[S_{NM}]^{\dagger}[I_{NL}]$ where $\dagger$ denotes transpose. We have chosen three substances which are red ink, stain and the base parchment ($M=3$). Since the base parchment is not perfectly homogeneous in terms of condition and thickness, so that it is expected that this technique can reveal the in-homogeneity by using a reference spectra from one pixel in 8249 pixels. Details of the component spatial pattern analysis method are described in references [4, 5, 6].

Figures 6 (a), (b), (c), (d) show the visible image, extracted mapping of parchment, red ink, and stain. As shown in Figure 6(b), red ink text was clearly distinguished from the base parchment. Two stain spots on the top recognised in the visible image were successfully removed. The stained area, on the other hand, appeared as shown in Figure 6(c). The intensity of the extracted image is probably due to the content of each substance. As shown in Figure 6(d), the parchment is not perfectly homogeneous, and blue lines in the visible image were also recognised. Assuming that the blue pigment is either lapis lazuli or azurite, no particular absorption peak appeared in the frequency range from 0.5 to 2 THz, and then it is difficult to distinguish blue
colour and the base parchment in this frequency range.

Experimental results reveal that THz spectroscopy and THz-imaging can distinguish inks and stain, and will help character recognition in a medieval manuscript. We are developing a new imaging system with THz-FTIR system which has wide frequency range up to around 15 THz to increase applicable art materials. Imaging system with quantum cascade lasers which are tuned at the absorption peak frequencies of substances, such as 3.8 THz for azurite, are desired as non-invasive on-site analysers By constructing a quantitative database of art materials, it is expected that THz spectroscopy can analyse pigment to binder ratio and mixture ratio of different pigments and/or binders.

5 Conclusion

We applied THz spectroscopy and THz-imaging with a component analysis method for text recognition of a medieval manuscript. The red ink text and stains on the text were successfully distinguished. Although further investigation is required to distinguish other colours and different types of stains on aged materials and to understand which molecular movement makes particular peaks in the spectra, we believe that THz spectroscopy can contribute to text recognition of historic manuscripts and can be used as a non-invasive analysis method for conservation science of cultural properties.

Acknowledgments

The authors thank Prof. Tomonori Matsushita of Senshu University for giving the opportunity to apply THz spectroscopy for character recognition research, and thank Mr. Richard A. Linenthal of Bernard Quaritch, London, for supplying the rare binding fragments from medieval manuscripts. The authors also thank to Dr. Chiko Otani of Riken for facilities at Riken and useful discussions.