반도체디스플레이기술학회지 제16권 제2호(2017년 6월)
Journal of the Semiconductor & Display Technology, Vol. 16, No. 2. June 2017.

테라헤르츠를 이용하여 글자를 읽어내기 위한 전처리 과정에 대한 연구

박인호[*] · 김성윤[**] · 김영섭[*†] · 이용환[***]
[*†]단국대학교 전자전기공학과, [**]단국대학교 전자공학과, [***]원광대학교 디지털콘텐츠공학과

A Study of the Use of Step by Processing for the Reading Letters Using Terahertz

Inho Park[*], Seongyoon Kim[**], Youngseop Kim[*†] and Yonghwan Lee[***]
[*†]Department of Electronic and Electronical Engineering, Dankook University
[**]Department of Electronic Engineering, Dankook University
[***]Department of Digital Contents, Wonkwang University

ABSTRACT

Recently, ancient documents are actively studied and discussed. However, ancient documents has a few problems on interpretation. The antique documents are too fragile to hand over. So, some studies have been carried out using terahertz to read ancient documents without damaging them. Three techniques are necessary to read letters using terahertz. First, PPEX algorithm, which distinguishes pages. Second, TGSI technique, which distinguishes text from paper on a page. Third, CCSC algorithm, which transforms signals to letters. In this paper, we will describe the preprocessing process to facilitate the recognition of letters before applying the post processing as we mentioned above. Histogram equalization, Histogram stretching and the Sobel filter were applied to the preprocessing.

Key Words : Terahertz, Histogram Equalization, Histogram Stretched, Sobel Filter

1. 서 론

최근 역사에 대한 인식이 중요해지면서 고문서 분석에 대한 연구도 많이 이루어지고 있다. 고문서의 경우 손만 대도 바스러지는 경우가 많고, 기존의 활용방식인 X-선의 경우 인체에 유해하므로, 이에 대한 대체 기술로 테라헤르츠(Terahertz)가 각광받고 있다.

테라헤르츠(Terahertz)는 투과성을 가진 전자파로써 10의 12제곱을 뜻하는 테라(Tera)와 진동수 단위인 헤르츠(hertz)를 합성한 용어이다. THz로 표시하며 테라헤르츠 방사선(terahertz radiation) 또는 줄여서 티-선(T-ray)이라고도 한다.

테라헤르츠를 이용하여 글자를 읽어내는 연구에서 가장 중요한 핵심은 페이지를 구분하는 것, 페이지에서 글자와 종이를 구분하는 것, 그리고 테라헤르츠파를 이용하여 받은 신호들을 변환하는 과정을 통해서 글자로 만들어 내는 것이 세 가지이다 [1].

본 논문에서는 테라헤르츠를 이용하여 글자를 읽어내는 연구에서 글자 구분에 정확도를 높이기 위한 전처리 과정에 대해 서술하고자 한다.

본 논문의 구성은 다음과 같다. 1장에서는 서론을, 2장에서는 테라헤르츠를 이용하여 글자를 읽어내기 위한 관련 기술을 서술하고, 3장에서는 제안하는 알고리즘을 설명하며, 4장에서는 실험결과를 보이고 5장에서 결론을 맺는다.

2. 관련기술

테라헤르츠를 이용하여 글자를 읽어내기 위한 알고리

[†]E-mail: wangcho@dankook.ac.kr

즘은 Fig. 1과 같다. 먼저, 테라헤르츠를 이용하여 촬영을 한다. 이후 페이지를 구분하는 단계, 글자의 잉크를 추적하는 단계, 글자를 인식하는 단계의 순서를 거쳐 사람의 눈을 통해 글자를 인식한다.

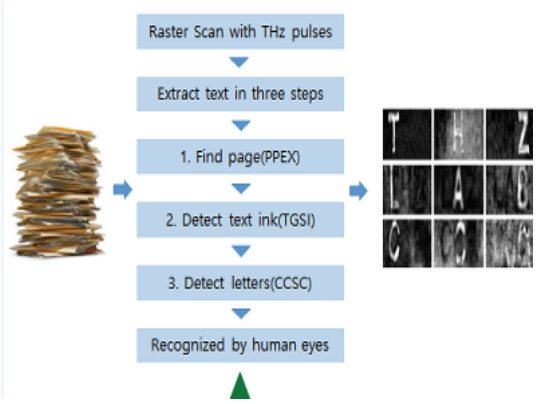

Fig. 1. Reading through a closed book algorithm.

PPEX (Probabilistic Pulse Extraction)알고리즘은 페이지를 구분하기 위해 적용되는 기술로, 짧은 간격을 두고 테라헤르츠파를 여러 차례 발생시킨 후 각 파장이 책장에 부딪혀 되돌아오는 시간을 측정하여 페이지를 구분하는 알고리즘이다. Fig. 2은 PPEX를 그림으로 표시한 것이다 [2,3].

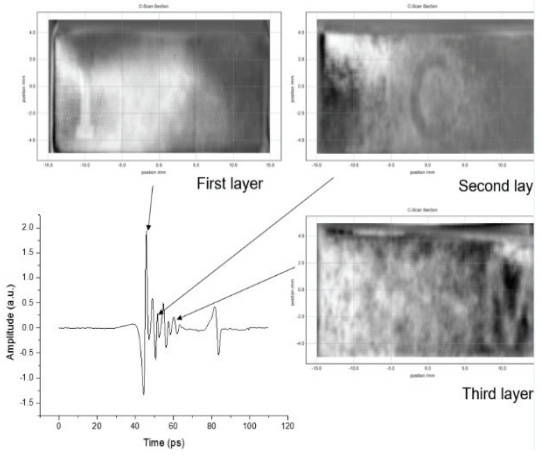

Fig. 2. PPEX (Probabilistic Pulse Extraction).

TGSI (Time-gated Spectral Imaging) 기술은 종이와 잉크의 전자파 흡수율이 다르다는 것을 이용한 기술로, 시간적 스펙트럼을 이용하여 잉크와 글자를 구분한다. Fig. 3은 TGSI를 보여준다. [4]

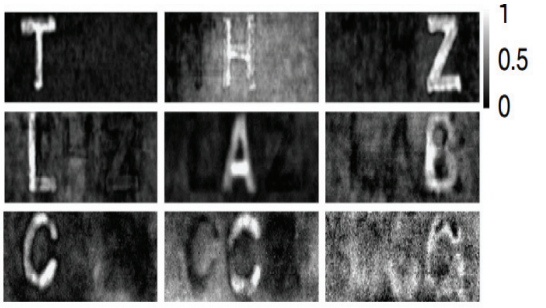

Fig. 3. TGSI (Time-gated Spectral Imaging).

CCSC (Convex Cardinal Shape Composition) 알고리즘은 잉크와 글자가 구분되어 잉크로 인식된 부분을 글자와 매칭하여 글자를 찾아내는 알고리즘으로, 인공지능 기술이 적용되어 있다. Fig. 4는 CCSC를 보여준다 [5,6].

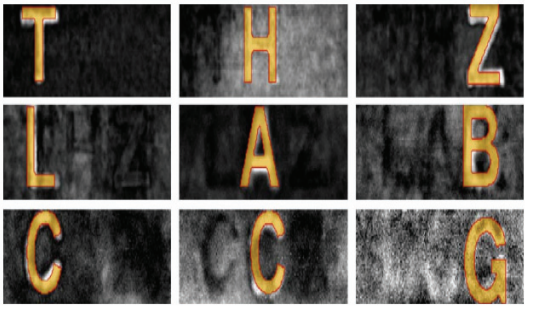

Fig. 4. CCSC (Convex Cardinal Shape Composition).

3. 제안하는 알고리즘

본 논문에서는 글자 인식에 대한 정확성을 높이기 위해서 전처리 과정을 추가한다. Fig. 5는 본 논문에서 제안하는 글자를 읽어내기 위한 전처리 과정이 포함 된 알고리즘이다. 전처리 과정에 사용되는 알고리즘에 대해서는 글자의 명암비 조정과 관련된 다양한 알고리즘을 적용해 보았다.

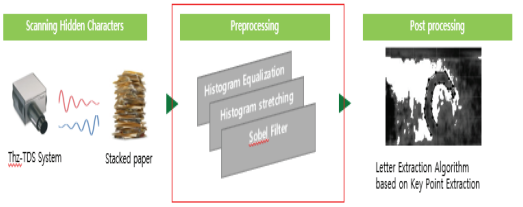

Fig. 5. Proposed algorithm.

4. 실험결과

본 실험은 본연구실에서 테라헤르츠 촬영을 통해 확보하고 있는 영상을 이용하였고, 실험환경으로 AMD A10-7700K 3.40GHz CPU 와 MATLAB 프로그램을 이용하여 histogram equalization(hist eq), histogram stretched(hist st), Sobel필터를 이용하였다 [7,8]. Fig. 6은 본 실험에 활용된 2 layer에 있는 알파벳 C 이미지이다.

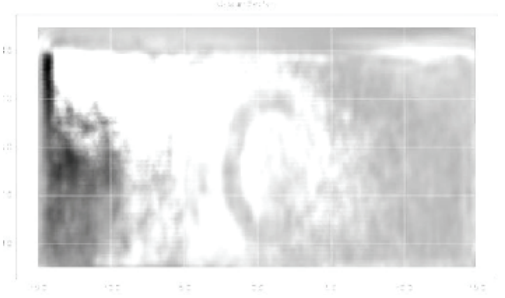

Fig. 6. Original Image (alphabet C / 2layer).

Fig. 7은 Sobel 필터를 이용한 결과물이다.

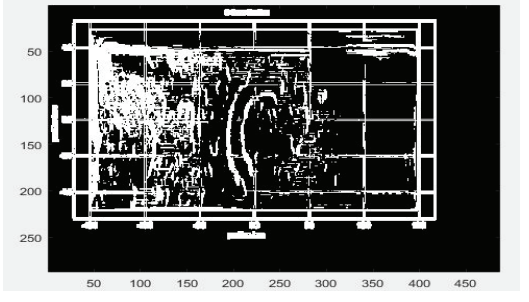

Fig. 7. Sobel filter(alphabet C / 2layer).

Fig. 8은 histogram equalization을 적용한 결과물이다.

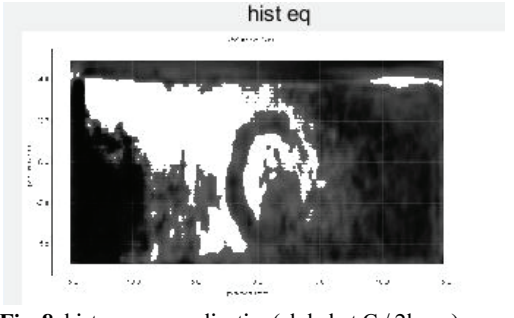

Fig. 8. histogram equalization(alphabet C / 2layer).

Fig. 9는 histogram stretched을 적용한 결과물이다.

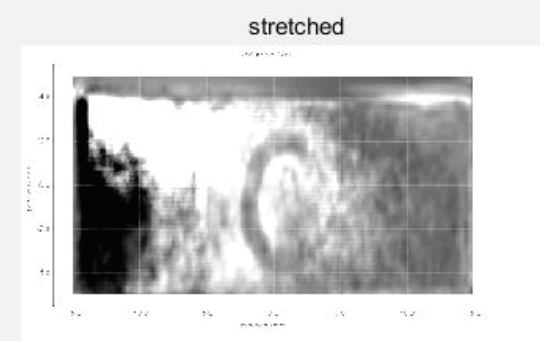

Fig. 9. histogram stretched (alphabet C / 2layer).

Fig. 10은 hist eq 적용 후 hist st을 추가로 적용한 결과물이다.

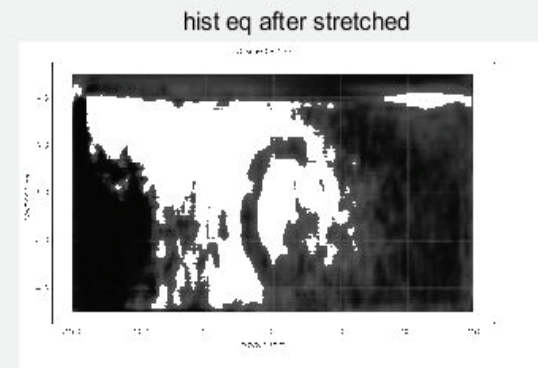

Fig. 10. histogram equalization after histogram stretched (alphabet C / 2layer).

Fig. 11은 hist st 적용 후 hist eq를 추가로 적용한 결과물이다.

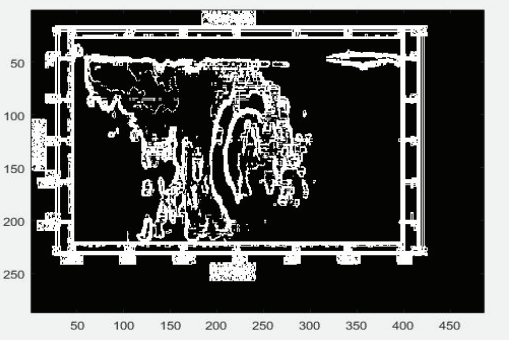

Fig. 11. histogram stretched after histogram equalization (alphabet C / 2layer).

Fig. 12는 hist st와 hist eq를 순차적용 후 Sobel을 추가로 적용한 결과물이다.

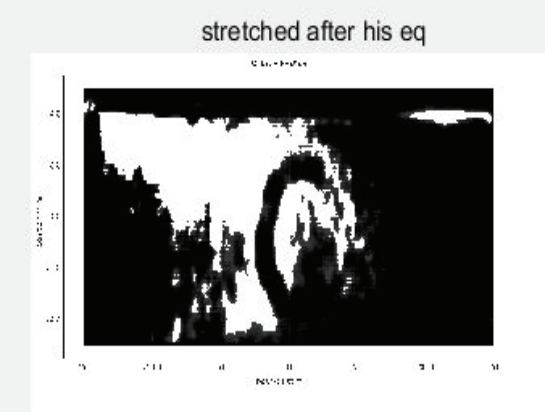

Fig. 12. histogram stretched after histogram equalization after Sobel filter (alphabet C / 2layer).

5. 결 론

본 논문에서는 테라헤르츠를 이용하여 글자를 읽어내는 과정에 대한 연구를 진행하면서 글자 인식에 대한 정확도를 높이기 위해 전처리 과정 이용하는 방법에 대해 실험하였다.

histogram stretched, histogram equalization, Sobel filter를 이용 및 순차적 적용한 실험 결과, histogram stretched을 이용한 방법과 histogram equalization을 적용한 방법이 가장 좋은 결과를 보였다. 본 연구는 현재까지 위에서 언급한 필터들을 적용하여 결과를 도출한 단계에 있다. 현재 적용한 필터 외에 다른 필터를 적용하고 조합하여 글자를 추출하는데 있어 더욱 효율적인 새로운 알고리즘을 개발하는데 초점을 두고 연구를 진행하고 있다.

감사의 글

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2017-000738, 덮여진 책의 글자를 읽어내는 연구 개발)

참고문헌

1. Redo-Sanchez,A., Heshmat, B.,Aghasi, A., Naqvi, S., Zhang, M., Romberg. J., Raskar. R., “Terahertz time-gated spectral imaging for content extraction through layered structures”, Nature Communications, 7, 12665, 2016.
2. Manceau, J.-M., Nevin, A., Fotakis, C. & Tzortzakis, S. “Terahertz time domain spectroscopy for the analysis of cultural heritage related materials”. Applied. Physics. B 90, pp. 365–368, 2008.
3. Fukunaga, K. & Hosako, I. “Innovative non-invasive analysis techniques for cultural heritage using terahertz technology”. Comtes Rendus Physique. 11, pp. 519–526, 2010.
4. Galva˜o, R., Hadjiloucas, S., Bowen, J. & Coelho, C. “Optimal discrimination and classification of THz spectra in the wavelet domain”. Optics. Express 11, pp. 1462–1473, 2003.
5. Aghasi, A., Romberg, J. “Convex cardinal shape composition”. SIAM J. Imaging Sciences. 8, pp. 2887–2950, 2015.
6. Aghasi, A., Romberg, J. “Learning shapes by convex composition”. arXiv preprint arXiv:1602.07613., 2016.
7. Sobel,I.E. Camera Models and Machine Perception, Ph.D. Thesis, Stanford University (1970). Dissertation
8. Rafael C. Gonzalez and Richard E. Woods, "Digital Image Processing", USA: Pearson Education, 2001.

접수일: 2017년 6월 22일, 심사일: 2017년 6월 24일, 게재확정일: 2017년 6월 24일