OCR 사용자 가이드

김동환

신민재

**목차**

**1. 해결하려고 하는 것**

**2. 사례조사**

**3. 해결 방법 제시**

**4. 제공 기능**

**5. 사용법 (Windows 기준)**

**1. 해결하려고 하는 것**

영어나 한글 이미지 파일을 TXT형식으로 글자 인식 데이터 추출

**2. 사례조사**

**Tesseract-ocr**

- 유니 코드 (UTF -8)을 지원하며 100개 이상의 언어를 바로 인식할 수 있다.

-Tesseract는 일반 텍스트, hOCR (HTML), PDF, 보이지 않는 텍스트 전용 PDF, TSV 등 다양한출력형식을 지원합니다. 마스터 브랜치는 또한 ALTO (XML) 출력을 실험적으로 지원한다.

-Tesseract **는 다른 언어를 인식하도록 훈련 될 수 있다.**

**- windows 나 linux같은 환경에서도 실행 할 수 있으며 무료 소프트웨어이다.**

**3. 해결 방법 제시**

**Tesseract-ocr의 경우 영어 기반의 OCR 프로그램으로 한글 인식에 굉장히 취약하고 변형된 문자 인식에 굉장히 취약하다.**

**새로운 OCR프로그램을 개발하기보다는 Tesseract는 품질 향상 및 다른 언어의 습득이 가능하여 이를 수정하고 향상하여 한글에 최적화 된 프로그램으로 향상시키겠다.**

**4. 제공 기능**

**한글이나 영어의 문자 텍스트 이미지를 입력하면 텍스트형식의 글자로 추출된다.**

**OCR 대상 글씨체**

**- 한글 및 영문에서 사용률이 매우 높은 폰트**

**굴림, 궁서, 돋움, 바탕, 맑은 고딕, 신명조, Arial, Times New Roman등**

**OCR 작업 대상**

**- 인쇄되어 발행된 매체의 스캔 이미지 파일**

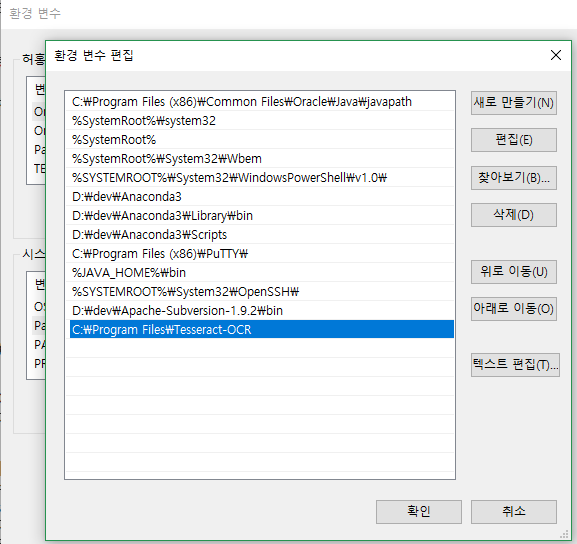
**신문(대부분 고딕 및 명조 글씨체 사용), 도서(다양한 글씨체 사용)**

**5. 사용법(Windows 기준)**

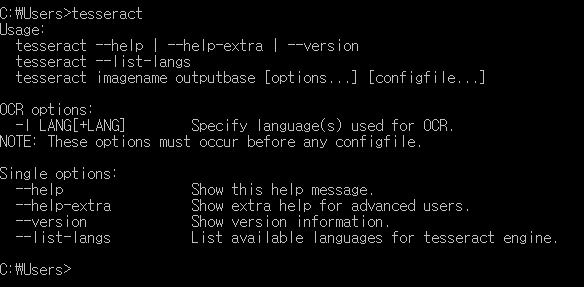
**1. tesseract-ocr 다운로드**

<https://tesseract-ocr.github.io/tessdoc/Home.html>의 사이트로가서 자신의 기준에 맞는 tesseract-ocr 버전을 다운로드한다.

2. 환경변수 설정



자신에 맞는 tesseract-ocr 경로를 환경 변수에 추가하여 준다.



**위와 같이 나오면 설치와 환경 변수가 완료된 것이다.**

**3. Viet-ocr다운로드**

Tesseract-ocr을 위한 JAVA/.net GUI frontend 이다.

Tesseract가 지원하는 여러 언어의 광학 문자 인식을 지원한다.

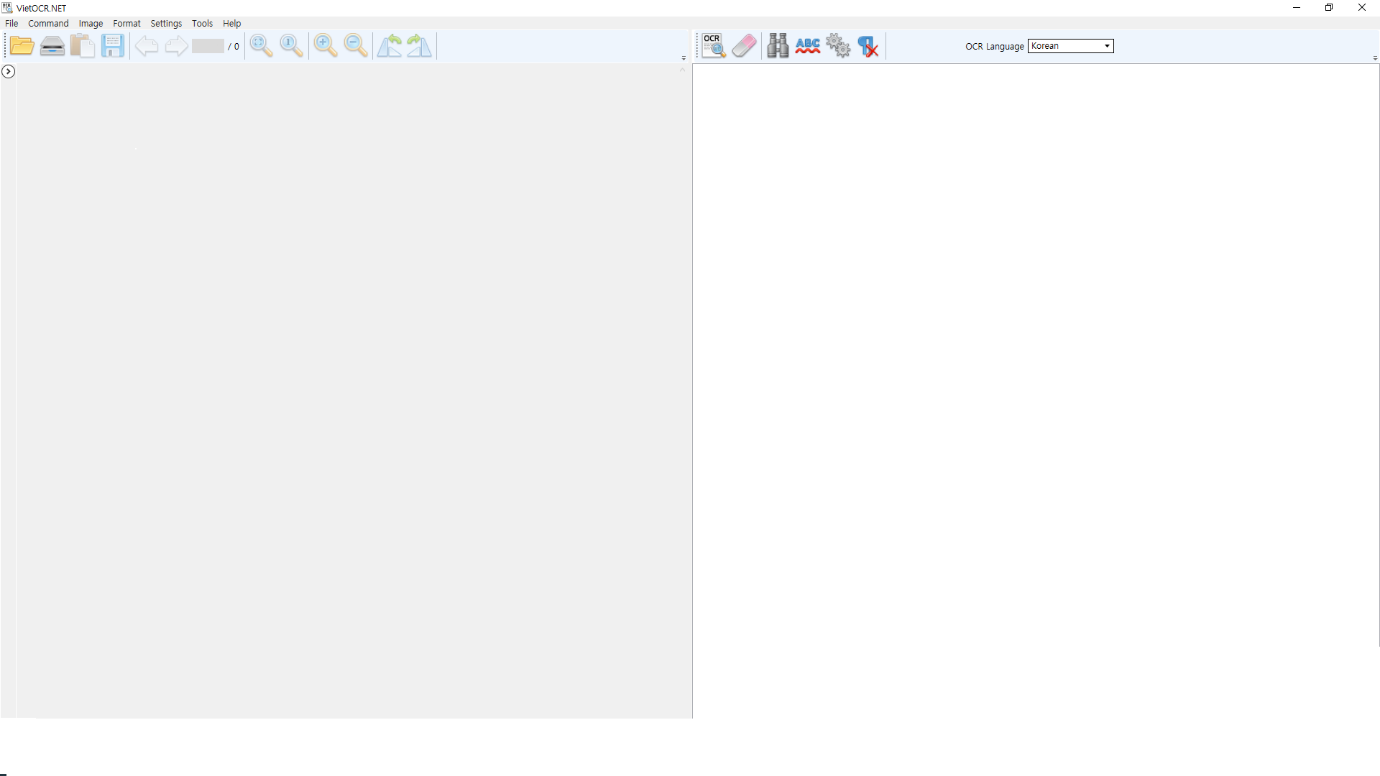
<https://sourceforge.net/projects/vietocr/> 사이트에서 자신에 맞는 버전을 다운로드한다.

4. <https://github.com/kimdong-hwan/o-s> 의 vs2008파일은 tesseract-OCR와 viet/tesseract-cor 폴더내에 압축 풀어준다.

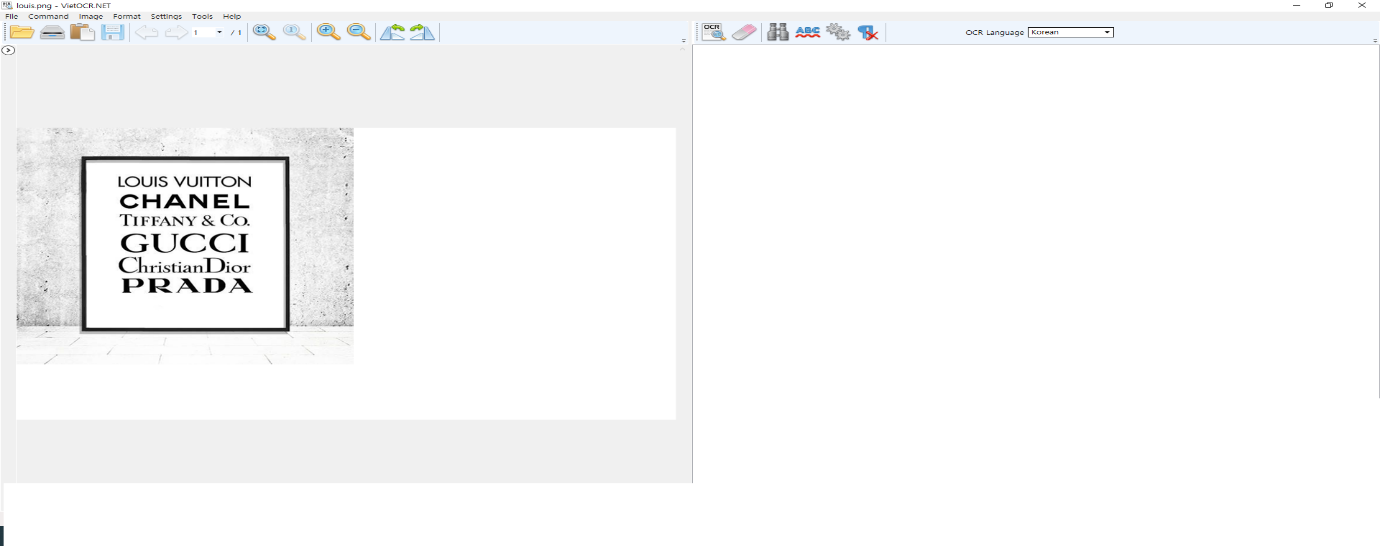
5. tessdata파일은 안에 있는 두 파일을 tessract-Ocr/tessdata와 viet/tessract-cor/ tessdata 파일 내에 풀어 기존 파일을 덮어 씌우면 된다.

6. vietocr실행

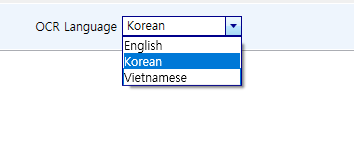
**기본 화면**

****

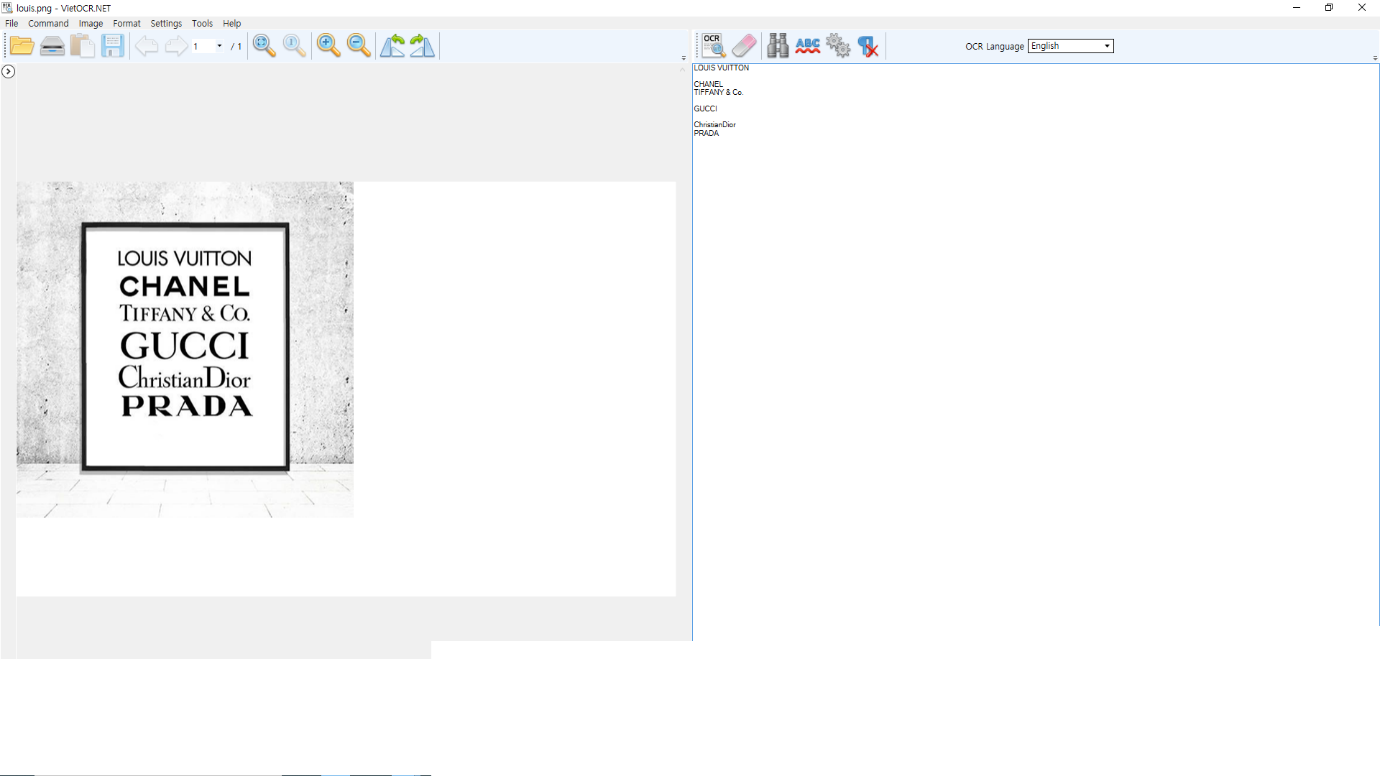
**왼쪽 위 메뉴에서 File-open 후 자신이 원하는 이미지파일을 선택한다.**

****

**오른쪽 위의 ocr languae에서 원하는 언어를 선택 후(원하는 경우 두가지 언어 동시 선택 가능)**

****

**오른쪽 창 위의 Perform OCR 버튼을 눌러준다**

****

**텍스트형식으로 인식이 완료된 상황이다.**

**팀 Github :** <https://github.com/kimdong-hwan/o-s>