Tesseract OCR 기반 인쇄 서적의 키워드 모니터링 시스템 설계

이주찬*, 김무중*, 유윤섭*
*한경대학교 전기전자제어공학과
e-mail: ysyu@hknu.ac.kr

Design of keyword monitoring system of printing paper based on Tesseract OCR

Ju-Chan Lee*, Mu-Joong Kim*, Yun-Seop Yu*
*Dept of Electrical, Electronic and Control Eng, Han-Kyong University

요 약

디지털 정보 처리 및 습득 속도에 대한 관심이 높아지면서 이와 관련된 많은 연구가 수행되고 있지만 아날로그 정보에 대한 연구는 많이 부족하다. 따라서 본 논문에서는 자동으로 책을 넘기고 각 페이지의 시잔을 촬영하여 컴퓨터로 전송한 후에 Tesserac-OCR을 이용하여 이를 디지털화 하여 저장하고 사용자가 원하는 키워드가 존재하는 페이지를 찾아 출력하는 시스템을 설계 및 구현한다.

1. 서론

21세기 정보화 시대에 접어들면서 대부분의 데이터가 디지털 형태로 저장되고 지식의 습득 또한 디지털 정보를 통해 이루어 질 것으로 예측하였다[1]. 하지만 현재까지 대부분의 연령층에 90% 이상이 전자 서적보다 인쇄 서적을 선호하는 경향이 나타났으며, 5년 후에 대한 설문조사에서 또한 60% 이상의 응답자가 여전히 인쇄 서적을 사용할 것이라는 결과가 조사된 바 있다[2]. 전자 서적의 경우에 정보가 디지털 형태로 존재하기 때문에 정보처리가 비교적 간편하며 빠른 검색 및 수정이 가능하다. 하지만 인쇄 서적의 경우 원하는 정보를 찾기 위해서는 사용자가 직접 책을 넘기며 정보를 찾아야 하고 수정이 거의 불가능 하다는 단점이 있을 뿐만아니라 이를 디지털 정보로 바꾸기 위해선 누군가가 직접 타이핑을 통해 이를 입력해야 하기 때문에 아날로그 정보의 디지털화 또한 쉽지 않다. 이러한 문제점을 해결하기 위해서 본 논문에서는 자동으로 책장을 넘기며 각 페이지의 사진을 촬영하고 촬영된 사진을 컴퓨터로 전송한 후에 이 정보를 구글에서 제공하는 Tesseract-OCR[3]을 이용해 글자를 데이터화 및 분석할 수 있는 시스템을 소개한다. 이 시스템을 통해 인쇄 서적의 정보 처리가 수월해 질뿐만 아니라 아날로그 정보의 디지털화를 가능하게 하여 4차 산업혁명의 핵심인 빅데이터의 정확도 및 질을 높일 수 있는 가이드 라인을 제공할 수 있을 것으로 예상한다.

2. 시스템 설계 및 구성

Ⅰ. 하드웨어 구성

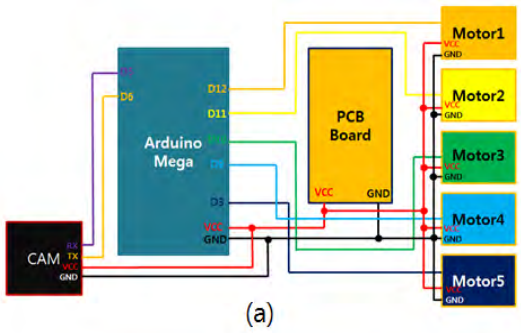

그림 1. (a) 하드웨어 블록도, (b) 작품 사진

그림1.(a)은 책을 자동으로 넘기기 위한 작품의 하드웨어 구성을 나타낸 것이며 그림. 1.(b)는 설계된 작품의 사진을 나타낸 것이다. 모터에는 서보모터 4개와 DC모터 1개가 사용되었으며, TTL 카메라를 사용하여 MCU로 사진을 전송하고, 이를 다시 컴퓨터로 전송하여 저장하도록 설

계하였다.

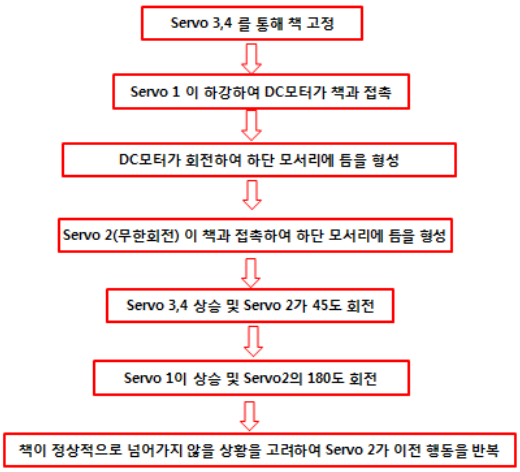

그림 2. 하드웨어 동작 알고리즘

그림2는 하드에어의 동작 알고리즘을 나타낸 것이다. 우선 Servo 3, 4가 하강하여 책을 고정한다. 이는 책과 DC 모터 사이에서 나타날 수 있는 마찰력을 증가시켜 2장씩 넘어가는 것을 방지 및 하단 모서리에서 나타나는 틈의 높이를 제어한다. 다음으로 Servo1이 하강한 후 DC모터가 회전하여 책과의 틈을 형성한다. 다음으로 Servo 3,4가 상승하고 Servo2가 45도 회전하여 생성된 틈으로 들어간다. 이때 Servo2가 180도 완전회전 할 경우 Servo1과 힘의 방향이 겹치기 때문에 책에 손상이 생길 것을 우려하여 45도 우선회전 하도록 설계하였다. 다음으로 Servo1이 상승한 후 Servo2가 완전 회전 한다. 통신 방법은 RXTX를 이용한 유선 통신을 사용하였다.

Ⅱ Tesseract-OCR

그림 3은 구글에서 제공하는 Tesseract-OCR의 구동 순서를 나타낸 것이다[3].

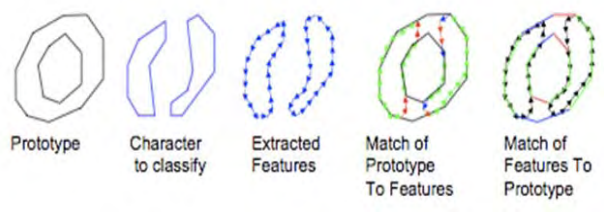

그림 3. Tesseract-OCR의 구동 순서

Tesseract은 특징점 추출과 MMSE 필터를 통한 판독을 통해 글자를 인식하는데, 이를 위해 우선 가우시안 필터를 적용하여 저주파 성분을 모두 제거하고 Edge 부분만을 추출한다. 이후에 특징점을 추출하여 점으로 나타내게 되고, 이를 표준화 시킨 후 MMSE 필터를 통해 글자의 일치 여부를 확인하게 된다. 일반적인 Tesseract-OCR의 경우 노이즈나 왜곡된 글자에서 정확도가 크게 감소하는 결과가 나타난다. 따라서 이를 최소화하기 위한 전처리 과정을 추가하였다. 우선 픽셀 수를 늘리는 방법을 사용하였는데, MMSE 필터를 적용할 때, 특징점의 수가 많을수록, 정확도는 증가하게 된다. 따라서 특징점의 수를 늘리기 위해 픽셀수를 증가시켰으나 그만큼 연산의 양이 증가하여 판독 시간이 증가하였다. 글자의 팽창 연산은 푸리에 영역에서 글자의 주파수가 바뀌는 지점 주위에 Mask를 생성하여 주변의 Mask 내 픽셀에 해당 픽셀값을 적용시키는 연산을 말한다. 글자를 팽창하였을 때, 특징점이 증가하여 평균 정확도는 상승하였으나, '0' 또는 'a'와 같이 유사 판독될 수 있는 글자의 경우에 정확도가 다소 감소하는 경향이 나타났다.

3. 결론

책을 자동으로 넘기며 각 페이지의 사진을 찍고 이를 컴퓨터로 전송하여 글자를 판독해 주는 시스템을 Tesseract-OCR을 기반으로 설계해 보았다. Servo 모터 4개와 DC모터를 사용하여 책을 넘기는 하드웨어를 구성하였으며, 통신은 RXTX 유선 통신을 사용하였다. 픽셀 수를 조절 및 팽창 연산을 추가하여 특징점 및 MMSE 필터 연산의 정확성을 상승시켰으며, 이때 연산수의 증가로 판독 시간이 크게 증가하였다.

참고문헌

[1] 서은영, 김성혁, 오경묵, "국가 지식정보자원의 디지털화 관리를 위한 전략", 정보관리학회지, vol.17 pp.213-234, 2000

[2] 박혜경, "전자책의 사회적 선택과 지식정보 생산양식의 변화", 사회과학연구, vol.21, pp.213-235, 2010

[3] Tesseract-OCR manual ver 4.0, Google International, 2017