컬러 영상에서 문자정보 추출을 위한 이진화 및 잡음 제거

정장호(한밭대학교 정보통신공학부)
이은주(한밭대학교 컴퓨터공학부)

요 약

오늘날 자료를 쉽게 정리하고 검색하기 위해 문서에 포함된 문자자료들을 데이터베이스화 하게 된다. DB화 하고자하는 자료가 컬러 영상의 표제영역일 경우, 스캐너 등의 입력장치로 입력받아 컬러 영상의 문자영역에 포함된 문자들을 인식하여야 한다.

현재 오프라인 문자인식(OCR)은 한글 및 영문자와 한자까지도 인식할 수 있는 알고리즘이 많이 개발되었다. 또, 제한된 분야에서는 98%정도의 인식률을 보이는 인식기가 개발되어 있으며, 인식률을 높이기 위한 새로운 알고리즘이 계속 개발되고 있다.

그러나 아직 배경이 컬러 영상인 문서에서는 배경과 문자가 혼합되어 있어, 배경과 문자의 분리가 어려워, 문자를 완전하게 인식할 수 있는 인식기가 없는 상태이다. 배경에 그림 및 컬러가 포함된 문서의 인식을 위해서는 컬러 영상에서 문자영역만을 추출해야 할 필요성이 있다. 이에 본 논문에서는 컬러 영상에서 문자정보 추출 및 잡음 제거 방법을 제안하였다.

기존 연구에서는, 주로 HSI모델의 컬러간의 관계성을 고려한 클러스터링을 통하여 문자영역과 배경영역을 분리한다. 그리고 분리된 문자영역의 잡음을 제거하기 위하여 외곽선 추적에 의한 라벨링을 하였다. 또는 HSI 각각의 히스토그램을 구하고 HSI에 대한 영상 내의 벡터들의 상관관계를 영상에 적용하여 적합한 임계값을 계산하고, 이 값에 의하여 컬러 영상에서 문자영역과 배경영역을 분리하였다.

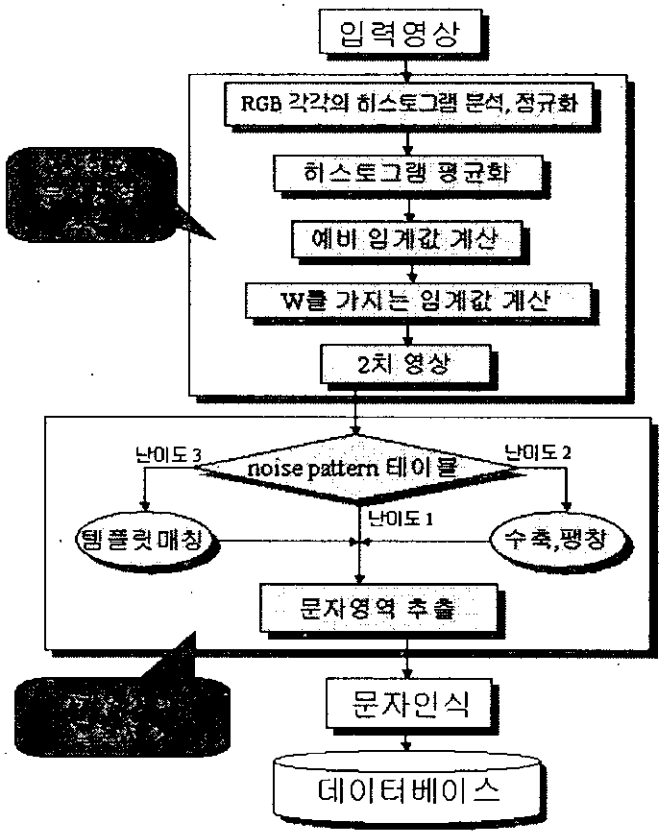
기존의 알고리즘의 경우 잡음량을 고려하지 않고 잡음을 제거하기 때문에 계산량과 시간이 많아진다. 또 라벨링에 의한 제거방법은 학술 논문지의 특성상 잡음의 픽셀크기가 작기 때문에 라벨의 갯수가 너무 많아져 계산량을 증가시킨다. 이에 본 논문에서는 잡음 제거시에 계산량과 시간을 줄일 수 있는 이진화 및 잡음제거에 관한 연구를 하였다.

입력 영상에 대하여, RGB모델을 기반으로 하여 영상의 히스토그램을 분석하여 정규화하고, 예비임계값을 구하며, 배경후보영역과 문자후보영역을 구한다. 배경후보영역과 문자후보영역간의 컬러 관계성 계수에 의하여 새로운 임계값을 가지고 이진화함으로 문자영역을 추출하였다. 히스토그램에서 배경영역과 문자영역을 분리하여 문자를 추출하기 위한 임계값 결정은 매우 중요하다. 단순히 히스토그램의 골짜기만을 가지고 임계값을 결정하게 되면 문자정보에 잡음이 많이 포함되거나, 반대로 문자정보가 소실되는 등 분리가 매끄럽지 않게 된다. 이에 본 논문에서는 배경후보영역과 문자후보영역의 컬러간의 관계성 계수를 정의하고, 계수가 크면 임계값을 크게 하여 문자영역의 손실을 최대한 줄인다. 또, 계수가 작으면 임계값을 작게 하여 배경영역에 포함에 되는 잡음들을 최대한 제거하고, 손상된 문자영역은 다시 팽창처리로 복원하는 방법을 제안하였다.

문자영역에 포함된 잡음의 제거는 이진영상을 분석하여 noise pattern 테이블을 만들고 그 분포도를 계산하여 난이도를 분류하였다. 분류된 난이도에 따라 처리 단계를 다르게 두어 계산량과 시간을 줄였다. 난이도 1은 잡음이 거의 없고 문자영역의 손상도 없어 잡음 제거 처리가 필요 없고, 난이도 2의 경우는 약간의 잡음이 포함되어 수축·팽창처리로 잡음을 제거하였다. 난이도 3의 잡

음을 갖는 영상은, 인식할 수 없는 정도의 많은 잡음이 포함되어 있는 경우로, noise pattern 테이블과 이진영상을 템플릿 매칭하여 잡음을 제거하였다.

배경에 그림 및 컬러가 포함된 영상에서, 제안한 알고리즘에 의해 추출된 문자와 일반 문서에 포함된 문자에 대한 인식 실험에서 대등한 수준의 인식률을 보여, 컬러 영상의 문자영역 추출에서, 본 방법의 유용성을 입증할 수 있었다. 향후 다양한 컬러 영상에 포함된 문자영역의 효율적 추출을 위하여, noise pattern 테이블 작성의 자동화에 대한 연구가 필요하다.



【제안한 방법의 흐름도】