OCR을 위한 문서 영상의 자동평가

윤병훈[O] 하진영
강원대학교 컴퓨터정보통신공학과
ysoftman[O]@kangwon.ac.kr, jyha@kangwon.ac.kr

Automatic Evaluation of Document Image for OCR

Byoung-Hoon Yoon[0] Jin-Young Ha
Department of Computer and Information Communications Engineering, Kangwon National University

요 약

본 논문에서는 OCR(Optical Character Recognition)의 정확도를 위해 인쇄체 한글 문서 영상에 대한 자동 평가방법을 제안한다. 자동 평가방법은 문서가 스캔된 상태에 따라 낮은 해상도, 영상 자체의 기울어짐, 많은 잡음 등을 판단하여 인식하지 않고도 인식률을 추측할 수 있다. 평가방법은 영상 자체의 밝기, 기울기, 영역의 특징, 문자의 상태 등을 특징 항목으로 만들어 점수를 산출한다. 각 항목의 점수는 가장 높은 인식률을 가지는 영상의 특징 값을 기준으로 삼는다. 각각의 특징에 대해 점수가 산출되면 인식률에 높은 비중을 차지하는 특징에 높은 가중치를 적용하여 최종 점수를 산출한다. 영상 평가방법을 통해 높은 점수를 얻은 영상은 상용 인식기를 통해 인식한 결과 높은 인식률을 나타냈고, 평가방법에서 낮은 점수를 받은 영상은 상대적으로 낮은 인식률을 나타냈다. 본 논문에서 제안하는 문서영상을 위한 자동 평가방법은 인식기를 사용하지 않고 영상의 품질을 측정하기 때문에 빠른 시간에 인식률을 추측할 수 있고, 낮은 인식률을 보일 수 있는 영상에 대해서는 항목별 점수를 피드백으로 사용할 수 있어 인식하기전 문서 영상의 전처리에 과정에 도움을 줄 수 있다.

1. 서 론

최근 인터넷 포털사이트들은 다양한 문서에 대해 검색 서비스를 제공하고 있지만 아직까지 그 기능은 사용자들의 기대에 미치지 못하고 있는 실정이다. PC가 보급되기 이전의 문서들은 종이에 기록되어 있기 때문에 이를 디지털화시키기 위해 문서를 스캔하여 영상을 인식하는 OCR(Optical Character Recognition) 기술을 사용해야 한다. 하지만 OCR을 사용하더라도 문서 보존 상태에 따라 변색되고 스캔 시 생기는 잡음 등의 영향으로 완벽한 상태의 영상을 얻을 수 없기 때문에 높은 인식률을 기대하기는 어렵다.

국내에서 최근 수년간 상용화한 문서 인식 시스템은 한글 인식에 초점을 맞추고 있으며, 속도 향상을 위해 한글 문자를 구조적인 특성을 이용하여 초성, 중성, 종성으로 분리한 후 자소 단위의 인식을 시도하는 것이 보편적인 경향이다[1]. 이러한 시스템에서는 영상 스캔 시 상대적으로 높은 해상도와 잡음이 없다는 것을 가정해야 하는 문제점이 있다. 따라서 원문서가 깨끗하지 않거나 스캔 시 문자의 크기와 스캐너의 해상도에 따라 인식이 불가능한 경우가 발생할 수 있다[2].

문서 자체의 상태와 스캔 시 영향으로 영상은 잡음과 해상도가 떨어지는 경우가 생긴다. 이런 영상을 인식하게 되면 인식자체가 불가능하거나 낮은 인식률을 나타낸다. 인식기가 높은 인식률을 산출하기 위해서는 영상 자체가 깨끗하고 왜곡이 없어야 한다.

본 논문에서는 인식기를 사용하지 않고 영상이 높은 인식률을 가질 수 있는지를 판단해주는 자동 문서영상 평가방법을 제안한다. 평가 방법은 스캔된 영상에 대해 특징을 추출하고 추출된 특징들을 평가하여 점수를 산출한다. 여기서 낮은 점수를 받은 영상은 실제 인식하였을 때도 낮은 인식률 나타냈고, 높은 점수를 받은 영상은 상대적으로 높은 인식률을 기록하였다. 자동 평가방법과 인식률의 상관정도를 파악하기 위해 인식률과 평가점수 간의 상관계수를 산출하였다.

2. 전처리 과정

기존의 스캔된 영상에 대해서 특징을 추출하여 평가하는 방법[3]은 영문자를 대상으로 글자영역에서 글자의 두께와 끊어진 정도 등을 기준으로 영상의 품질을 좋고 나쁨(Good or Poor)으로 판단한다. 하지만 복잡한 문서나 문서전체를 평가하기에는 특징항목이 부족하여 신뢰하기가 어렵다. 문서 영상을 인식하기 위해서는 잡음제거, 기울기 교정 등의 전처리과정이 선행되어야 한다. 표 1은 인식을 위한 대표적인 전처리 항목을 나타내고 있다.

표 1. 인식 전 전처리 요소

BINARIZATION	영상의 이진화 정도
NOISE	영상의 잡음 정도
SKEW	영상의 기울기 정도
ZONE	영상의 영역분류 정도
SEGMENTATION	영상의 문자분할 정도

- BINARIZATION

인식을 하기위해선 먼저 컬러(RGB)영상을 그레이스케일(Gray-Scale)영상으로 변환하고, 이를 다시 글자를 나타내는 검정색과 글자가 아닌 배경을 나타내는 흰색으로 구성된 이진(Binary)영상으로 변환하여야 한다.

0~255 단계로 구성된 그레이 영상을 이진영상으로 변환하는 방법은 기본적으로 영상 전체의 평균을 문턱 값(Threshold)으로 사용하는 GlobalThresholding이 있다. 하지만 이진영상은 글자색과 배경색이 뚜렷하게 구별되어야 하기 때문에 GlobalThresholding의 고정적인 문턱값만으로는 항상 좋은 결과는 얻기는 힘들다[4]. 이와 같은 문제점을 개선하기 위해 이지미의 글자색과 배경색을 클러스터링 해나가면서 유동적인 Threshold를 찾아가는 FBC(Foreground and Background Clustering) 방법을 사용한다[5].

- NOISE

문서 영상의 글자분분에 잡음이 섞이면 전혀 다른 글자로 인식되는 경우가 생긴다. 이러한 잡음이 글자의 한 부분인지 판단하여 글자부분이 아니라면 잡음을 제거해야 한다. 잡음제거를 위해선 크게 저주파통과필터(Low Pass Filter), 미디언 필터(Median Filter), Smoothing등이 있다. 이들 잡음제거 방법 중 원본 영상의 강한 Edge와 상세 한 부분을 보존하기위해서는 미디언 필터가 효과적이다.

- SKEW

문서 영상의 전처리는 스캐닝 과정에서 문서를 잘못 놓거나 자동 급지 장치(ADF: Automatic Document Feeder)의 다양한 속도 변이에 의해 발생하는 문서 영상의 기울어짐을 추정하고 교정하는 과정을 말한다. 문서 영상의 기울어짐은 영상의 스캐닝 과정에서 흔히 발생하는 문제다[6]. 문서 영상 기울기로 인해 왜곡이 심해지면 문서인식 자체가 불가능하기 때문에 기울이기 교정이 필요하다. 그림 1은 기울어진(Skew) 영상과 교정을 통한 올바른 영상을 나타낸다.

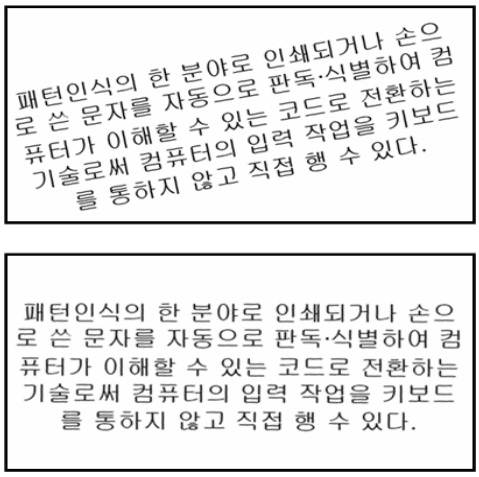

그림 1. 기울어진 영상과 교정영상

문서 영상을 인식하기 위해서 기울어진(Skew) 영상을 올바르게 되돌려야 한다. 문서 영상을 올바르게 되돌리려면 현재 문서의 기울어진 정도를 각도도 판단해야 한다. 이 각도는 보통 문서의 위쪽부분에 기울어진 정도를 판단하는데 이는 문서 전체에 대한 기울기로 활용되기에 부족하기 때문에 문서의 위, 아래, 왼쪽, 오른쪽 부분에서 각도를 검출한 다음 이들 각도들의 평균을 취하는 것이 좋다. 하지만 이 때 4부분의 각도들 중 나머지 각도들과 차이가 큰 각도는 불확실한 각도로 판단하고 제외시켜야 한다.

- ZONE

문서 영상 전처리를 통해 문서 영상의 기울어짐이 교정된 후에 문서 영상을 의미 있는 영역들로 분할하게 되는데, 각 영역들은 텍스트 또는 그래픽 정보(그림, 테이블, 선 등)를 포함한다. 이것은 문서 영상으로부터 기하학적인 구조를 추출하여 문서를 다양한 영역의 집합으로 분할하는 단계와 각 영역의 특성을 조사하여 텍스트와 비텍스트 영역으로 분류하는 영역 분류 단계로 구성된다[7]. 영역분류는 일반적으로 영상의 가로방향과 세로방향으로 투영하여 프로파일을 생성하고 이를 기준으로 영역을 나눈다. 그림 2는 투영 프로파일을 사용한 영상 영역분류 결과를 나타낸다.

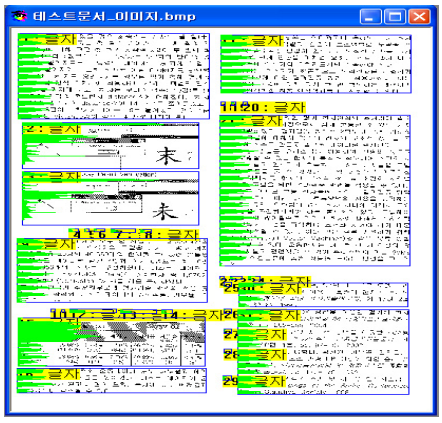

그림 2. 2단 문서 영상 영역 설정 결과

- SEGMENTATION

문서 구조 분석을 통해 추출된 텍스트 영역은 인식의 대상인 문자 단위로 분할하게 되는데, 일반적으로 투영 프로파일(Projection Profile) 방법을 적용하거나 연결요소 분석(connected component analysis) 방법 등을 적용한다[7]. 그림 3은 문자단위 분할 결과를 보여준다.

그림 3. 문자단위 분할 결과

3. 평가 방법

평가 방법은 우선 원본 영상을 입력으로 받아 평가항목에 해당하는 특징을 추출한다. 문서 영상 평가를 위해 2장에서 설명한 전처리 항목들에 대해 각각 점수를 산출한다. 각 평가항목들 점수는 다음과 같은 방법으로 영상에 대한 점수를 산출한다.

(1) Binarization Iteration(BI)

문서의 변색이나 짙은 배경색으로 글자와 배경의 구분이 어려운 영상은 좋은 이진 영상을 얻기 힘들다. 2장에서 설명한 FBC는 글자와 배경 화소들을 클러스터링 해가면서 각각 평균값을 찾아 나간다. 변색이나 짙은 배경색의 영상을 FBC방법으로 변환할 경우 클러스터링 과정이 일반 영상에 비해 많아진다. 영상의 이진화 평가는 클러스터링 과정의 반복 횟수로 판단하여 점수를 산출한다. 그림 4는 클러스터링 반복 횟수를 반환하는 의사코드(Pseudo Code)를 나타낸다. 먼저 영상전체의 평균값을 구하고, 이를 기준으로 흑화소와 백화소를 분류한다. 그리고 흑화소로 분류된 화소들에서의 평균값과, 백화소로 분류된 화소들에서의 평균값으로 다시 흑백화소를 분류한다. 이러한 과정을 각각의 평균값이 변하지 않을 때까지 반복 실행한다.

```
for (y=0; y<Height; y++)
  for (x=0; x<Width; x++)
    if (abs(Pixel[y][x]-mean_F) < abs(pixel[y][x]-mean_B))
        sum_F=sum_F+ptr[y][x];
        cnt_F++;
    else
        sum_B=sum_B+ptr[y][x];
        cnt_B++;
  if (cnt_F>=0&&cnt_B>=1&&sum_F>=1&&sum_B>=1)
    if ((mean_F!=sum_F/cnt_F)&&(mean_B!=sum_B/cnt_B))
        mean_F=sum_F/cnt_F;
        mean_B=sum_B/cnt_B;
    else    return iteration;
```

그림 4. FBC 클러스터링 과정 반복 횟수 반환

(2) Noise Ratio(NR)

원본 영상의 잡은 화소 제거를 위해 미디어 필터를 적용한다. 하지만 미디언 필터를 사용하더라도 완벽한 잡음제거는 할 수 없다. 이 때문에 잡음평가 항목에서는 미디어 필터를 적용한 영상에서 잡음으로 판단된 화소들의 비율을 판단하여 점수를 산출한다. 잡음의 비율(NR : Noise Ratio)은 다음과 같이 나타낼 수 있다.

$$NR = \frac{NoisePixels}{(WidthPixels \times HeightPixels)}$$

(3) Skew

정확하게 스캔된 문서 영상은 기울기가 0도에 가깝다. 하지만 각도가 1도만 기울어도 문자의 왜곡이 심해지기 때문에 ±5도의 범위를 벗어난 영상은 스캔이 잘못된 것이라고 판단하여 0점수를 산출한다. 기본적으로 각도는 시계방향이 양수 값을 나타내도록 한다. 그림 5는 문서 영상의 기울기 허용 범위를 나타낸다.

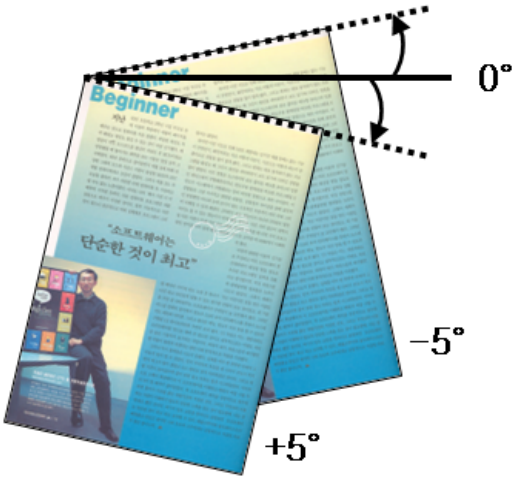

그림 5. 허용 범위의 각도

정확한 영상의 각도를 파악하기 위해 영상을 4부분(Top, Left, Bottom, Right)으로 나누고 각 부분의 각도를 측정한다. 이 때 ±5도 범위를 벗어나거나 나머지 값들과 큰 차이를 보이는 각도가 있으면 제외시키고 평균 각도를 산출한다. 영상의 기울기 평가에서는 0도를 기준으로 ±5도의 범위에서 0.5도씩 점수를 차등 산출한다. 그림 6은 영상의 4부분에서 각도를 추출하는 예를 보여준다.

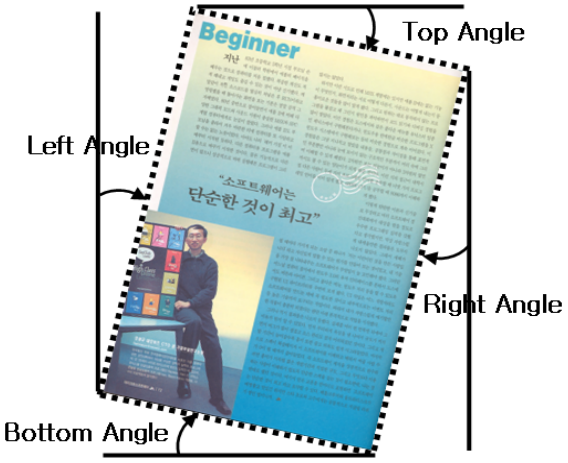

그림 6. 4 부분의 각도 추출

(4) Zone Distribution(ZD)

투영 프로파일을 사용하여 문서영상의 영역분류를 할 경우 그림영역은 하나의 영역으로 설정되어야 하고, 텍스트 영역은 보통 하나의 문단이 영역으로 설정되어야 한다. 하지만 불규칙한 글자 간격과 줄 간격 등으로 인해 설정된 영역이 너무 작거나 큰 경우가 생긴다. 그림 7은 분류 결과 너무 작게 설정된 텍스영역의 예를 보여준다.

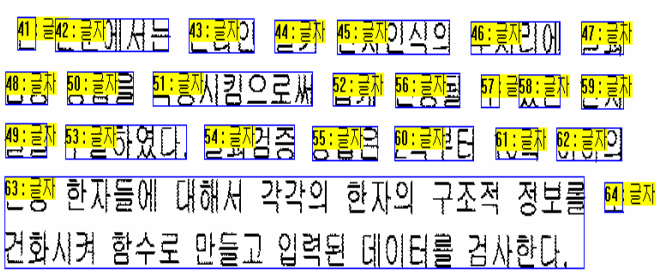

그림 7. 작게 설정된 텍스트 영역

영역분류를 통한 평가에서는 문서영상의 분류 결과 설정된 영역의 폭과 높이가 기준 이하인 영역들이 전체 영역에서 차지하는 비율을 점수로 환산한다.

(5) Segmentation Difficulty(SD)

영역분할에서 글자 영역으로 분할된 영역은 인식기가 인식할 수 있도록 문자분할 과정을 거쳐야 한다. 문자분할은 전처리 과정의 최종과정으로 인식에 가장 많은 영향을 미친다. 비정상적인 문자분할로 인해 인식결과가 전혀 다르게 나올 수 있다. 한글 문자분할의 경우 분할된 사각형(Rectangle)은 폭과 높이의 비율이 비슷해야 한다. 그림 8은 글자 간격이 너무 좁아 문자들이 합쳐진 비정상적인 문자분할을 보여준다. '스'와 '피'가 합쳐진 경우 분할 사각형의 높이와 폭의 길이가 상당한 차이를 보인다는 것을 알 수 있다.

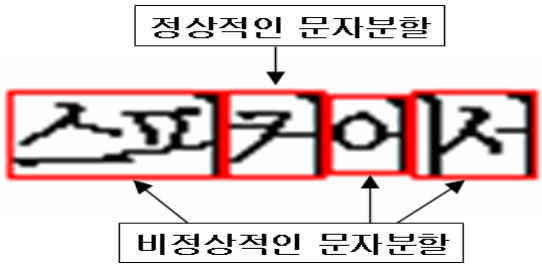

그림 8. 비정상적인 문자분할

문자분할 평가에서는 영역분류에서 텍스트로 설정된 영역들에 대해 문자분할을 시도하여 비정상적으로 분할된 문자들을 파악한다. 그리고 전체 문자분할에 대한 비정상적인 문자분할의 비율을 계산하여 점수를 산출한다.

각 평가 항목들의 점수들이 나오면 인식에 중요한 영향을 미치는 순서에 따라 가중치를 부여하여 100점 만점의 종합 점수를 최종적으로 산출한다. 문서 영상의 최종 평가 점수는 다음과 같이 산출된 각 항목의 점수에 가중치를 두고 계산한다.

$$Total_Score = w1(BI) + w2(NR) + w3(Skew) + w4(ZD) + w5(SD)$$

그림 9는 영상 평가방법 시스템의 전체적인 흐름을 나타내고 있다.

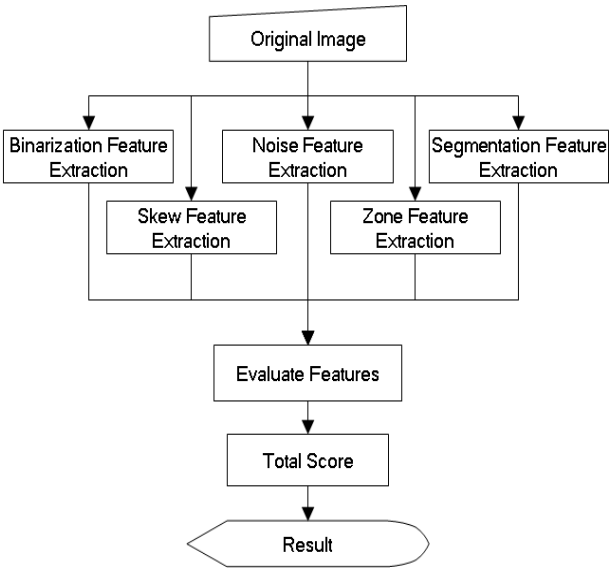

그림 9. 영상 평가 방법 흐름도

4. 실 험

문서영상의 자동 평가방법 실험을 위해 동화, 소설, 잡지, 만화 등의 다양한 문서에서 50개의 영상을 수집하였다. 수집한 영상의 인식률은 상용 인식제품인 아르미를 사용하였고 텍스트영역에서의 단어만을 인식하여 산출하였다.

자동 평가방법은 각 항목별 점수에 가중치를 부여하여 최종 평가점수를 산출한다. 이때 문자분할 평가항목은 전처리과정의 최종단계에 해당하여 기본적으로 다른 항목에 비해 높은 가중치를 두고 잡음 평가항목은 상대적으로 낮은 가중치 값을 지정한다. 표 2는 가중치를 각각 다르게 설정하여 4가지 경우를 나타낸다.

표 2. 평가항목별 가중치

CASE1	CASE2	CASE3	CASE4
w1=0.1	w1=0.1	w1=0.1	w1=0.2
w2=0.1	w2=0.1	w2=0.1	w2=0.1
w3=0.3	w3=0.2	w3=0.1	w3=0.1
w4=0.1	w4=0.1	w4=0.1	w4=0.1
w5=0.4	w5=0.5	w5=0.6	w5=0.5

자동 평가방법의 실험은 입력된 문서 영상을 5가지 항목으로 나누어 평가하여 항목별 점수를 산출하고 표 2에의 가중치 중 하나를 선택하여 최종 평가 점수를 산출하였다. 그림 10은 본 논문에서 제안하는 자동 평가방법을 사용하여 하나의 문서 영상을 평가한 결과화면이다.

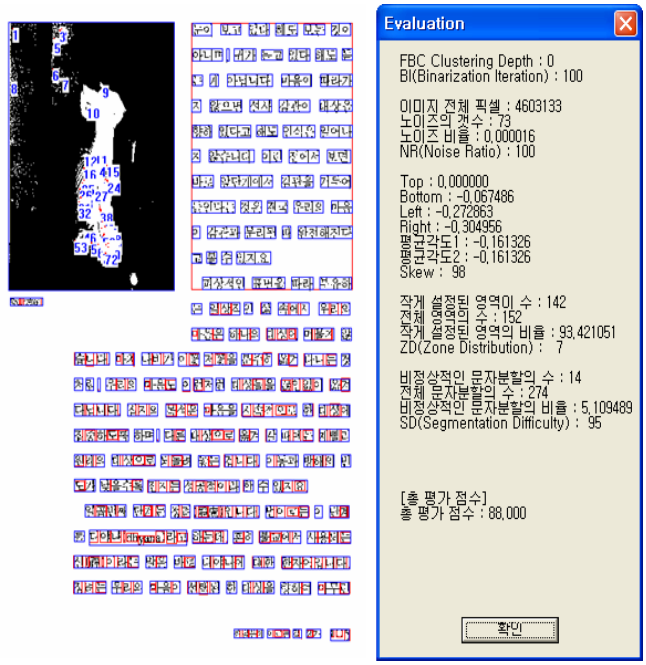

그림 10. 영상 평가 결과 화면

그림 10과 같은 평가방법을 모든 실험대상(50개문서영상)에 적용하여 각각의 평가 점수를 산출하였다. 그리고 4가지 경우의 가중치를 적용한 문서영상의 인식률과 자동 평가방법의 피어슨 상관계수(Pearson Correlation Coefficients)[8][9]를 파악하였다. 두 집단(X, Y) 간의 피어슨 상관계수(r)는 다음과 같다.

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}, \quad -1 \le r \le 1$$

각각의 가중치 적용 결과 첫 번째 가중치 설정(CASE1)이 인식률과 가장 높은 상관계수를 나타냈다. 표 3은 각각의 가중치에 대한 피어슨 상관계수 값을 나타낸다.

표 3. 가중치에 따른 피어슨 상관계수

	CASE1	CASE2	CASE3	CASE4
상관계수	**0.805**	0.784	0.741	0.758

가중치별 상관계수 실험결과 CASE1이 0.805로 가장 높은 상관계수를 보였는데, 이는 인식률에 가장 많은 영향을 미치는 요소가 첫 번째로 문자 분할과정이고 다음으로 문서영상의 기울기 교정과정이라는 것을 나타낸다.

5. 결 론

본 논문에서는 인식하기전 문서 영상에 대한 평가방법을 제안하였다. 자동 평가방법은 인식 전 주요 전처리 과정을 평가항목으로 만들고, 각 항목에 가중치를 부여하여 문서영상에 대한 최종적인 평가를 점수로 산출한다. 항목별 가중치는 실험을 통해 결정하여 문서영상에 대해서 자동평가실험을 하였다. 그 결과 평가방법에서 산출된 점수는 실제 인식률과 0.805의 피어슨 상관계수를 나타냈다. 상관계수를 통해 평가방법을 사용하여 높은 점수를 얻는 영상은 실제로 인식을 하였을 때도 높은 인식률을 보인다는 것을 알 수 있다.

문서 자동 평가방법은 인식을 하지 않고 영상의 인식률을 추측할 수 있기 때문에 평가방법의 각 항목별 결과를 피드백으로 사용하면 인식할 때 방해가 되는 요소를 미리 판단할 수 있어 최종 인식결과에 도움을 줄 수 있다.

실험에서 인식률을 구하기 위해 사용된 상용 아르미 인식기는 본 논문에서 제안한 자동 평가방법의 전처리 과정과 다소 차이가 있었다. 향후연구과제로 평가방법의 각 항목을 세분화하여 다양한 인식기에 적용 가능하도록 하는 연구가 필요하다.

6. 참고문헌

[1] 도정인, “한글 문서 인식 시스템의 개발”, 정보과학회지, 제 9권 제 1호, pp.22-32, 1991.

[2] 하진영, “통계적 방법과 구조적 정보의 결합을 통한 인쇄체 한글인식”, 강원대학교 정보통신 논문지, 제 2 집, pp.85-90, 1998.

[3] Luis R. Blando, Junichi Kanai, and Thomas A. Nartker "Prediction of OCR Accuracy Using Simple Image Features", Information Science Research Institute University of Nevada, Las Vegas, USA, IEEE, pp.319-322, 1995.

[4] J. He, Q.D.M Do, A.C. Downton and J.H.Kim "A Comparison of Binarization Methods for Historical Archive Documents", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition(ICDAR'05), IEEE, pp.538 - 542, 2005.

[5] Andreas E. Savakis, "Adaptive Document Image Thresholding Using Foreground and Background Clustering", Proceedings of international Conference on Image Processing(ICIP'98), IEEE, pp.785-789, 1998.

[6] 곽희규, “문서 영상의 단어 단위 분할 및 단어 영상의 속성 추출에 관한 연구”, 전남대학교 박사학위논문, 2001.

[7] 오일석, 김수형, 유태웅, 곽희규, “문서 영상 처리 기술과 디지털 도서관”, 정보과학회지, 제 20권 제 8호, pp.24-34, 2002.

[8] 박지선, 김택헌, 류영석, 양성봉, "추천 시스템을 위한 2-way 협동적 필터링 방법을 이용한 예측 알고리즘", 정보과학회논문지, Vol.29, No.10, pp. 669-675, 2002.

[9] Daniel Billsus, Michael J. Pazzani, "Learning Collaborative Information Filters," Proceedings of ICML, pp.46-53, 1998.