한국산학기술학회논문지
Vol. 10, No. 8, pp. 1869-1876, 2009

스캐닝 과정에서 발생하는 전자문서의 기하학적 변형감지에 관한 연구

오동열[1*], 오해석[2], 류성열[1]
[1]숭실대학교 IT대학 컴퓨터학과, [2]경원대학교 IT대학 컴퓨터 공학

Study on Measuring Geometrical Modification of Document Image in Scanning Process

Dong-Yeol Oh[1*], Hae-Seok Oh[2] and Sung-Yul Rhew[1]
[1]**Department of Computer Science, SoongSil University**
[2]**Department of Computer Science, Kyungwon University**

요 약 기존 종이 문서를 영상 파일로 변환하기 위해서 스캐너와 같은 광학기를 사용하게 된다. 스캐닝 과정에서 문서가 제대로 문서 영상으로 생성되었는지를 판단하기 위해서 이미지 품질 검사 과정을 거치게 된다. 이미지 품질 검사 과정에서는 스캐너 기기의 특성상 스캐닝 과정에서 발생할 수 있는 문서 영상의 기울기, 노이즈 여부, 문서가 접힌 상태로 스캔되었는지의 여부 등을 체크하게 된다. 이에 본 논문에서는 스캐너를 이용하여 문서 영상을 생성 과정에서 발생하는 기하학적 변형을 평가하기 위한 방법론을 제시한다. 본 연구에서는 품질 검사의 검사 항목에 대해서 영상 처리를 이용하여 각각의 변형 정도를 측정하고 각각의 변형 정도가 실제 문서의 가독성에 얼마나 영향을 미치는지를 OCR 결과 값과 비교한다. OCR 인식 성공 비율과 각 항목별 변형 정도를 나타내는 측정 값 간에 상관관계를 분석하기 위해서 피어슨 상관 계수(Pearson Correlation Coefficient)를 이용하고 이를 기반으로 실제 문서 영상의 변형 정도를 평가하기 위한 가중치 값을 산정한다. 제시한 방법으로 평가에서 높은 평가 값으로 계산된 영상 문서는 OCR 인식률에서도 높은 인식 결과를 나타내고 있다.

Abstract Scanner which is a kind of optical devices is used to convert paper documents into document image files. The assessment of scanned document image is performed to check if there are any modification on document image files in scanning process. In assessment of scanned documents, user checks the degree of skew, noise, folded state and etc This paper proposed to how to measure geometrical modifications of document image in scanning process. In this study, we check the degree of modification in document image file by image processing and we compare the evaluation value which means the degree of modification in each items with OCR success ratio in a document image file. To analyse the correlation between OCR success ratio and the evaluation value which means the degree of modification in each items, we apply Pearson Correlation Coefficient and calculate weight value for each items to score total evaluation value of image modification degrees on a image file. The document image which has high rating score by proposed method also has high OCR success ratio.

Key Words : Quality Assurance, Document Image Quality Check, Document Image Processing

1. 서론

일반적으로 전자적으로 유통되는 전자 문서의 경우, 생성 방식에 따라 다음과 같이 두 가지로 나뉘게 된다. 첫째는 전자 문서 자체가 전자적으로 작성된 경우로 워드나 PDF와 같이 컴퓨터를 이용하여 문서를 작성하고 파

*교신저자 : 오해석(oh@kyungwon.ac.kr)
접수일 09년 03월 20일 수정일(1차 09년 07월 02일, 2차 09년 07월 22일, 3차 09년 08월 06일) 게재확정일 09년 08월 19일

일을 생성하는 경우를 의미하며 컴퓨터에서 쉽게 내용 검색이나 자료의 저장 및 관리가 가능하다. 두 번째는 기존의 오프라인 종이 문서를 스캐너나 디지털 카메라와 같은 광학기기를 이용하여 이미지 파일로 생성하는 경우로 종이 문서의 상태나 스캐너에서 문서 생성시의 이미지 생성 옵션 혹은 물리적인 스캔 방식에 따라서 동일한 종이 문서의 경우에도 다양한 형태로 이미지가 생성되거나 혹은 이미지 자체에 변형이 발생할 수 있다. 실제 공인 전자 문서 보관소에 저장할 문서를 생성하는 전자화 시스템의 경우, 생성된 이미지 파일에 대한 품질은 사용자가 육안으로 직접 확인하여 이를 평가하게 되는데 생성된 이미지 파일 개수가 적은 경우에는 문제가 되지 않으나 이미지 파일의 개수가 많은 경우에는 많은 인력과 시간이 소요되게 된다. 이에 본 논문에서는 스캐너를 이용하여 전자문서를 생성하는 경우, 기기의 물리적인 특성에 따라서 발생할 수 있는 문서 이미지의 기하학적 변형에 대해서 유형별로 살펴보고 각각의 변형을 컴퓨터가 자동으로 인식할 수 있는 방법에 대해서 연구한다. 연구된 문서 이미지 변형 유형을 기반으로 각 변형 항목에 대해서 평가하고, 평가된 항목의 각 유형별로 서로 다른 가중치를 부여하는 방법에 대해서 논하고 생성된 문서의 재 스캔 여부를 통보하는 시스템의 설계 및 구현을 통해서 본 연구에 대한 효용성을 평가한다.

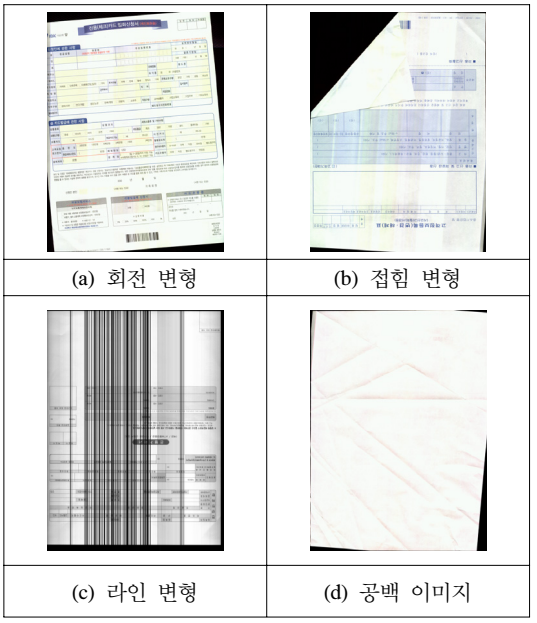
2. 관련 연구

기존 문서 이미지의 품질 검사는 OCR(Optical Character Recognition)의 문자 인식율과 관련하여 실제 문자 인식 단계 이전에 이미지의 내용 품질 검사를 통해서 문자 인식율에 적합한 이미지인지 아닌지를 평가하기 위한 연구로 진행되었다. Michael은 OCR 에러 비율을 미리 예측하고 이를 평가하기 위한 평가 알고리즘을 제시하였다[1]. Michael은 평가 알고리즘을 통해서 서로 다른 문서 이미지에 대해서 문자 영역에 존재하는 문자 인접 화소(Touching Characters)와 문서 내 흰색 화소(White Speckle) 및 문서 외 검정 화소(Black Speckle)를 추출하여 이를 OCR 에러 비율을 측정하기 위한 평가 값으로 제시하였다. Michael은 실제 영상 이미지의 평가를 위해 평가된 예측 값과 실제 OCR 에러 비율을 비교하여 제시한 세 가지 항목에 대한 효용성을 평가하였다. Henry는 문서 이미지내의 문자 영역을 추출하여 회전 각도(Skew Angle), 문자의 명료도(Blur Kernel), 이진화 설정값(Threshold), 민감도(Sensitivity), X 비율(X scaling), Y 비율(Y scaling)의 6가지를 문서 이미지의 품질 평가를 위한 평가 값으로 제시하였다[2]. Andrea는 문서 이미지의 품질을 평가한 후, 문자 인식 비율을 높이기 위한 필터를 자동으로 선택하는 방법에 대하여 연구하였다[3]. 제안 연구에서는 이미지의 품질 평가를 위해서 폰트의 크기(Font Size), 수직/수평 선의 두께(Stroke Thickness Factor), 문자 인접 화소(Touching Character Factor), 문자 외 작은 크기 화소(Small Speckle Factor), 문자 내 흰색 화소(White Speckle Factor)와 같이 6가지 항목을 문서 이미지의 품질을 평가하기위한 값으로 제시하였다. 이를 기반으로 각 평가 값에 적합한 최종 모폴로지 연산을 통해서 문서 이미지에 적합한 필터를 적용하였다. Ben M. Chen은 디지털 카메라로 획득된 문서 영상의 왜곡에 대해서 보정 방법을 제시하면서 문서 영상에 존재하는 문자의 경계선(Character Boundary)와 문자의 최상단 꼭지점(Tip Point)를 문서 영상의 왜곡 측정을 위한 특징으로 추출하였다[4]. Wei Dong은 문자가 존재하는 문서가 아닌 영상 파일에서의 영상에 대한 평가 방법을 제안하였다[5]. Wei Dong은 영상 파일내의 픽셀을 에지 부분(Edge Region), 텍스쳐 부분(Texture Region)과 평활화 부분(Flat Region)의 세 가지 부분을 영상 품질 평가를 위한 평가 값으로 제시하였으며, 각 블록 별로 러프 퍼지 정수(Rough Fuzzy Integral)을 적용하여 영상 이미지의 품질을 검사한다. 이와 같이 기존 연구에서는 실제 대상 문서가 올바르게 광학기기를 통해서 디지털화 되었다는 가정에서 영상 이미지에 존재하는 문자 부분을 추출하여 문자 인식에 적합성 연구를 판단하는 형태로 진행되었다. 이에 본 논문에서는 종이 문서가 스캔되는 과정에서 발생하는 여러 가지 기하학적인 변형에 대해서 생성된 이미지를 분석하여 문서의 재 스캔 여부를 평가하는 시스템을 설계하고 개발한다.

3. 스캔된 문서 이미지의 변형

이미지 생성을 위한 스캐너는 문서의 급지 방식에 따라서 플랫베드(Flat Bed)형과 ADF(Ahead Document Feeder)형으로 구분된다. 본 논문에서는 다량의 종이 문서 스캔에 적합한 ADF형 스캐너를 이용하여 종이 문서를 스캔 하는 경우에 발생할 수 있는 이미지 변형 유형을 다음과 같이 네 가지로 제시한다. 첫째, 그림 1의 (a)와 같이 문서 스캔 과정에서 발생하는 이미지의 회전이다. 이미지의 회전은 실제 용지가 피더(Feeder)에서 삽입되는 경우, 용지가 수직으로 적재 되지 않은 상황에서 발생하거나 용지를 흡입하기 위한 롤러의 마모로 인해서 발생되는 왜곡이다. 둘째, 그림 1의 (b)와 같이 용지 스캔 시

용지가 접힘으로서 발생되는 왜곡이 있다. ADF형 스캐너의 경우, 이중 급지(Double Feed)를 검사하지만, 이중 급지 센서가 전 구간에 걸쳐서 존재하지 않기 때문에 모서리가 접힌 상태나 파기된 상태에서 이를 감지하지 못하고 종이 문서가 스캔될 수 있다. 셋째, 그림 1의 (c)와 같이 수직 혹은 수평으로 라인이 생기는 경우이다. 일반적으로 스캐너의 이미지 센서의 경우 CIS(Contact Image Sensor)를 사용하게 되는데, CIS 자체는 실제 스캔 되는 매체와 접촉을 통해서 전자 이미지를 생성하는 방식으로 스캐너 내부에 스테이플과 같은 이물질이 존재하는 경우에는 용지가 스캔되는 과정에서 수직 선분의 노이즈가 발생할 수 있다.



(a) 회전 변형 (b) 접힘 변형

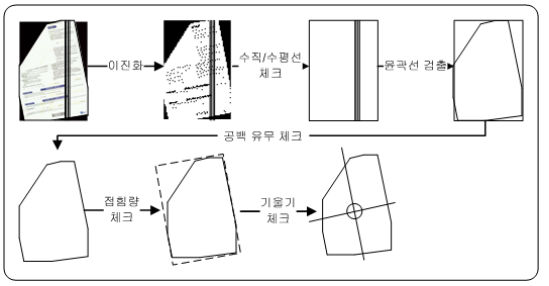
(c) 라인 변형 (d) 공백 이미지

[그림 1] 스캔 문서의 변형 유형

넷째는 공백 이미지 스캔으로 일종의 변형이라기보다는 실제 공백 문서가 스캔되는 경우가 있다. 이와 같은 경우에는 그림 1의 (d)와 같이 모든 픽셀이 흰색 값을 갖지 않고, 접힘 혹은 이물질에 의해서 특정 농도 이하의 픽셀 값을 가지게 된다. 마지막으로 기존에 언급한 회전 변형, 접힘 변형, 라인 변형, 공백 이미지 스캔이 복합적으로 발생되는 경우가 있다.

4. 이미지 변형의 감지

본 논문에서 제안하는 이미지 변형의 감지 방법은 그림 2와 같다. 첫째, 이미지 이진화 과정을 거쳐 모든 픽셀을 흑백으로 재 정의한다. 둘째, 생성된 이미지의 수직 및 수평 투영을 통해서 이미지 전 구간에 걸쳐 발생하는 수직/수평선을 파악한다. 셋째, 문서의 바깥을 둘러쌓고 있는 상하좌우에 위치한 윤곽선을 추출한다. 넷째, 추출된 윤곽선내의 각 픽셀별 흑화소 분포도를 계산하여 공백 유무를 검사한다. 공백 유무 체크 시에는 수직/수평선 유무를 파악하여 윤곽선내 흑화소인지 수직/수평선인지를 고려해야 한다.



[그림 2] 단계별 이미지 변형의 감지

다섯째, 윤곽선내의 전체 면적을 구하고 이를 이상적인 이미지 문서의 면접비와 비교하여 접힘 여부를 파악한다. 마지막으로 문서가 공백문서인 경우에는 문서의 외곽선을 기준으로 문서의 기울기를 검출하고 문서가 공백문서가 아닌 경우에는 이미지 내 문자열을 추출하여 그 기울기를 검출한다.

4.1 문서 영상의 이진화

이진화 방법은 이진화를 위한 임계치의 설정 방법에 따라 전역적 이진화와 지역적 이진화 방법으로 분류된다. Sezgin는 다양한 이진화 방법에 대해서 문서 인식의 관점에서 비교 평가를 수행하였다[6]. 전역적 이진화로 Otsu의 방법이 빠른 속도와 문서 이미지의 이진화에 적합한 것으로 연구되었다. 이에 본 논문에서는 문서 영상의 이진화를 위하여 Otsu가 제안한 이진화 방법을 적용하여 문서 영상을 이진 변환을 위한 임계치 값을 결정한다[7].

4.2 문서 영상의 수직/수평선 변형

CIS의 노이즈에 발생하는 문서 영상의 수직/수평선은 이미지의 높이 및 너비 전 구간에 걸쳐 직선 형태로 발생되는 선분들로 표현된다. 문서 영상의 수직/수평선을 검출하기 위해서 너비가 M이고 높이가 N인 문서 이미지에 대하여 다음과 같이 수직/수평 프로파일링(P_h/P_v)이 실행한다.

$$P_h(y) = \sum_{i=0}^{M} Img(i,y) \quad (1)$$

$$P_v(x) = \sum_{i=0}^{N} Img(x,i) \quad (2)$$

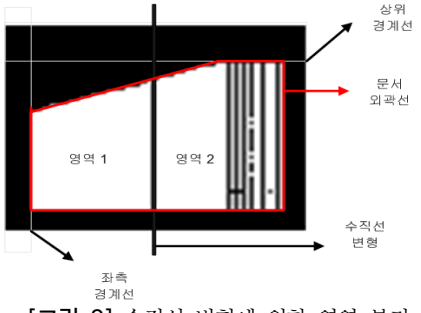
이진화 과정을 거쳐서 이진영상으로 변환된 각 픽셀의 값은 0(흰색)과 1(검정색)으로 표현되기 때문에 이미지에 수평/수직선이 존재하는 경우에는 해당 P_h의 원소 중에서 수평선이 존재하는 경우에는 해당 값이 문서 영상의 너비 W와 동일하고 P_v의 원소 중에서 수직선이 존재하는 경우, 해당 값이 문서 영상의 높이 H와 동일하게 된다. 따라서 실제 수직/수평선 변형으로 인해 생성된 선분의 면적(Dim_h/Dim_v)은 다음과 같다.

$$Dim_h = \sum_{i=0}^{M} \mathrm{if}\, P_h(i) = w\,?\,w : 0 \quad (3)$$

$$Dim_v = \sum_{i=0}^{N} \mathrm{if}\, P_v(i) = h\,?\,h : 0 \quad (4)$$

4.3 외곽선 추출 및 문서 내 면적 계산

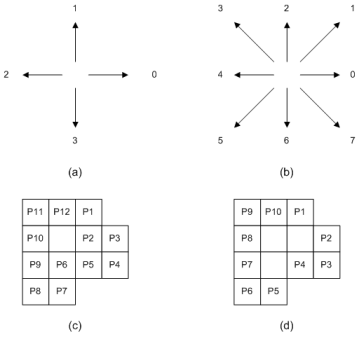
일반적으로 물체 혹은 영역의 경계선을 추출하기 위해 영상을 2차원적 위상에 근거하여 체인 코드(Chain Code) 방식을 사용한다. 이미지 내에 수직/수평 변형이 존재하는 경우에 기존의 4방향 체인코드나 8방향 체인코드를 적용하여 사물의 외곽선을 분류하게 되면 그림 3과 같이 실제 추출하고자 하는 외곽선은 하나의 문서 영역 내인데 비해서 수직/수평선으로 인해 여러 개의 영역으로 분리된다. 이러한 문제를 해결하기 위해서 기존의 8-방향 체인코드에 방향성 필터(Directional Filtering)를 적용한 S. Hoque의 방법론을 확장하여 문서의 진행 단위를 결정하는 방법을 제시한다[8]. 본 논문에서는 기존 문서 영상의 수직/수평선에서 검출된 수직선 중에서 너비가 최대인 값을 체인코드의 검색 단위로 사용을 한다.



[그림 3] 수직선 변형에 의한 영역 분리

4.4 공백 이미지 체크

추출된 문서 이미지 내에 존재하는 잡음을 제거하기 위해서 문서 영역 내에서 3X3의 폐쇄(Closing) 연산을 수행한다[9]. 폐쇄 연산 이후, 그림 4의 (b)와 같이 8방향 체인코드를 이용하여 문서 이미지 외곽선 내에 존재하는 대상들을 추출 한다.



[그림 4] 4/8 방향 체인코드

외곽선 내에 픽셀에 대하여 8방향 체인코드를 통해서 추출된 영역 경계의 면적을 계산하여 6폰트 이하인 체인코드 집합(Set)에 대해서는 노이즈로 간주하고 6폰트 이사인 체인코드 집합이 하나라도 존재하는 경우, 공백 이미지가 아닌 것으로 간주하고 더 이상의 외곽선 추출을 진행하지 않는다.

4.5 문서 영상의 접힘양 검출

본 논문에서 언급하고 있는 접힘양은 실제 문서의 일부분이 접힌 상태로 스캔이 되는 경우뿐만 아니라, 종이 문서가 스캐너의 스캔 대상 설정 폭 이상으로 회전되어 피딩(Feeding)되는 경우에 발생되는 일부 이미지의 유실도 이에 해당한다. 문서의 접힘양은 4.3에서 추출된 외곽선내의 영역 R의 면적으로 다음과 같다.

$$Area(R) = \int_x \int_y f(x,y)\,dydx \approx \sum_x \sum_y f(x,y) \quad (5)$$

여기서 화소가 영역 R의 내부에 있으면 f(x, y)=1이고 그렇지 않으면 0이다. 위와 같은 방식으로 실제 외곽선내의 영역이 계산되면 스캔 원 문서의 이상적인 영역 I의 면적과 비교하게 된다. 이상적인 영역 I란 문서가 접힘 없이 올바르게 스캔된 경우이며 문서의 너비가 W이고 높이가 H인 경우 다음과 같다.

$$Area(\dot{I}) = \int_{x}^{w}\int_{y}^{h} 1\,dxdy \approx \sum_{x}^{w}\sum_{y}^{h} 1 \qquad (6)$$

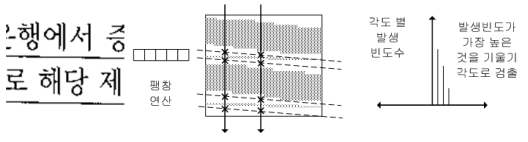
R의 면적과 I의 면적 차를 I의 면적과 다음과 같이 비교하여 접힘 비율 R(F)를 산출한다.

$$Ratio(F) = \frac{Area(\dot{I}) - Area(R)}{Area(\dot{I})} \times 100 \quad (7)$$

4.6 문서 영상의 이미지 기울기 검출

문서 영상의 기울기를 검출하기 위한 방법으로 공백이 아닌 문서에서 문자열을 추출하고 이를 기준으로 그림 5와 같이 이미지의 기울기를 검출하는 방법이 사용된다. 검출된 문서는 Hough의 변환(Hough Transform)을 이용하여 추출하는 방식이 많이 사용되나 많은 연산시간과 메모리를 요구한다[10]. 본 논문에서는 다음과 같은 방법으로 행간 중심점을 특징으로 문서의 기울기를 검출한다.

① (2W+1) 마스크를 사용하여 수평적으로 팽창 연산을 수행
② 일정 너비 기반의 수직 투영을 통한 행간 중심점 추출
③ 행간 중심점을 이용하여 기울기 검출
④ 기울기 각도별 히스토그램 분석을 통해서 가장 빈도수가 많은 각도를 문서의 기울기로 사용



[그림 5] 문서 영상의 기울기 검출

문서 영상의 경우, ±10° 이상이 기울어진 경우에는 문서의 재 스캔이 필요하다는 판단에서 0점으로 간주하고 -10°에서 10°도까지 산출된 각도에 가중치 값을 부여하여 100점 만점으로 기울어진 각도에 대한 차를 평가 값으로 환산한다.

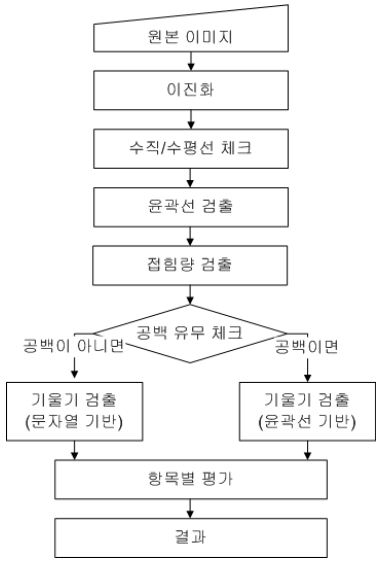
5. 문서 영상 품질 평가

본 5장에서는 문서 영상의 품질 평가 방법에 대하여 기술하고 실험 환경과 그 결과에 대해서 언급한다.

5.1 문서 영상의 품질 평가 방법

스캔 과정을 통해서 생성된 이미지에 대한 품질 평가는 [그림 6]과 같다. 첫째, 체크된 수직/수평선의 면적은 전체 이미지 대비 수직 수평선 면적의 비율을 계산하고 이를 전체 면적을 100을 기준으로 하고 수직/수평선 면적의 차를 평가 값으로 환산한다. 둘째, 접힘양은 실제 윤곽선 검출을 통해서 측정된 문서의 영역과 실제 스캔되는 문서가 변형이 발생되지 않았을 경우에 면적 비율로 이 또한 전체 면적을 100을 기준으로 하여 유실된 접힘면과의 차를 평가 값으로 한다. 셋째, 측정된 문서의 기울기는 ±10° 이상인 경우에는 0점으로 환산하고 그 사이각인 경우에는 측정된 실제 각도에서 가중치 10을 곱하여 0도인 경우 100을 만점으로 기울어진 각도에 가충치 값을 곱한 값의 차로 평가 값을 환산한다. 평가된 각 항목은 가중치 값을 두고 다음과 같이 계산한다.

$$Total\,Score = w_{FM} \times FM + w_{LM} \times LM + w_{SM} \times SM$$
FM은 접힘 변형 평가 값, LM은 라인 변형 평가 값,
SM은 기울기 변형 평가 값을 의미하며,
단 $w_{FM} + w_{LM} + w_{SM} = 1$ (8)



[그림 6] 단계별 이미지 품질 평가

5.2 품질 평가를 위한 가중치 값 선택

품질 평가를 세 가지 측정 항목인 회전 변형(SM), 접힘 변형(FM), 라인 변형(LM)중에 문서 영상의 OCR 인식률과 가장 연관이 있는 가중치를 적용하기 위해서 각 항

목별로 평가된 점수와 문서 영상의 OCR 인식률 간에 상관 관계를 파악한다[1-4, 13]. 두 개의 변수 간에 어느 정도 강한 관계가 있는지를 분석하는 단순 상관 분석(Simple Correlation Analysis)의 경우 피어슨 상관계수(Pearson Correlation Coefficient), 스피어만 상관 계수(Spearman Correlation Coefficient)와 크론바흐 상관 계수(Cronbach Alpha) 신뢰도가 있다. 스피어만 상관계수의 경우, 자료의 값 대신 순위를 이용하는 상관계수이며[11], 크론바흐 상관 계수 신뢰도의 경우 한 검사 내에서 변수들 간에 검사문항들이 동질적인 요소로 구성되어 있는지를 분석하는 것이다. 본 논문에서는 항목별 평가 점수와 OCR 인식율의 상관 관계를 파악하기 위해서 피어슨 상관 계수(Pearson Correl ation Coefficient)를 이용한다. 피어슨 상관계수는 두 변수 간의 연관 강도용 측정 지표로서 1과 -1 사이의 값을 가지며 1은 완전히 연관되었음을 의미하고, 0은 연관이 전혀 없음을 의미하며 -1은 완전 반대로 연관되었음을 의미한다[12]. 임의의 두 집단 (X, Y)에 간의 상관 관계를 나타내는 피어슨 상관계수는(r)는 다음과 같이 계산한다.

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 (Y_i - \overline{Y})^2}}, -1 \le r \le 1 \quad (9)$$

항목별 가중치 산출 방법은 표 1과 같이 각 항목별로 산출된 평가 값과 문자 인식율 간에 피어슨 상관계수를 도출 후, 피어슨 상관 계수의 총합에 각 항목별 피어슨 상관계수의 비율을 평가 항목별 가중치로 사용한다.

[표 1] 품질 평가를 위한 가중치 값 산출 방법

비교 대상		피어슨 상관 계수	항목별 가중치
평가 값 (회전 변형)	문자 인식율	rSM	$\frac{r_{SM}}{r_{SM}+r_{LM}+r_{FM}}$
평가 값 (라인 변형)	문자 인식율	rLM	$\frac{r_{LM}}{r_{SM}+r_{LM}+r_{FM}}$
평가 값 (접힘 변형)	문자 인식율	rFM	$\frac{r_{FM}}{r_{SM}+r_{LM}+r_{FM}}$

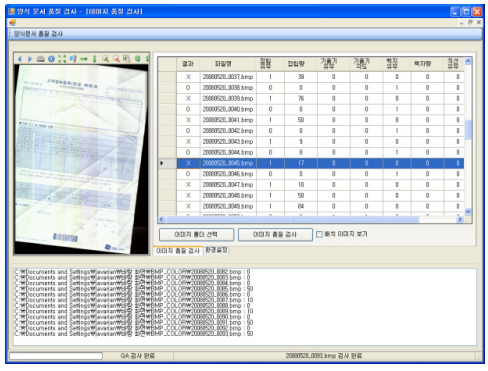
5.3 실험 및 평가

제안하는 문서 영상의 품질 측정을 위해서 문서 영상 품질 검사 프로그램은 VB.NET 개발언어로 구현하였으며 영상 처리를 위한 라이브러리로는 LeadTools을 이용하였다. 실험을 위해 A4형태의 양식 문서 100매를 서로 다른 ADF형 스캐너 세 종류를 이용하여 총 300매의 영상 문서 파일을 생성하였다. 수집한 문서 이미지의 인식률은 상용 인식 제품인 아르미를 사용하여 평가한다.

상용 인식기를 이용하여 문자를 인식 하는 경우, 각 문서 영상별 인식율은 다음과 같다. T는 문서 내에 존재하는 총 문자수를 의미하며 M은 오인식된 문자의 수, U는 미인식된 문자의 수를 나타낸다.

$$OCRRatio = \frac{M+U}{T} \times 100 \quad (10)$$

그림 7은 스캐너에서 생성된 문서 영상 300개에 대해서 각 문서 영상별로 접힘양, 공백 여부, 수직/수평선 및 기울기를 측정한 결과이다. 평가를 위해서 각 항목별 피어슨 상관계수를 기반으로 산출된 가중치 값은 표 2와 같이 접힘 변형에 경우, OCR 인식율과 가장 높은 상관 관계를 나타내고 있으며 라인 변형의 경우, OCR 인식율과 가장 낮은 상관 관계를 나타내고 있다.



[그림 7] 문서 영상 품질 측정 결과 예시화면

그림 8은 300개의 대상 영상 문서 중에 그림 7의 문서 영상 품질 측정 프로그램을 이용하여 공식 (10)을 적용한 결과이다.

[표 2] 300개의 영상 문서를 대상으로 산출된 가중치

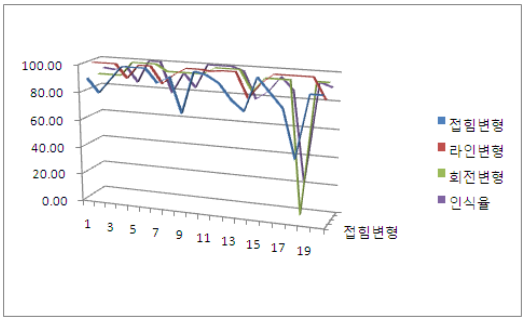
비교 대상		피어슨 상관 계수	항목별 가중치
평가 값 (회전 변형)	문자 인식율	0.93	0.49
평가 값 (라인 변형)	문자 인식율	0.16	0.08
평가 값 (접힘 변형)	문자 인식율	0.82	0.43

$$TotalScore = 0.43 \times FM + 0.14 \times LM + 0.53 \times SM \quad (10)$$

문서 영상의 품질을 객관적으로 평가하기 위하여 OCR 인식기를 이용한 문자 인식율과 본 논문에서 문서 영상의 품질 평가를 위해서 제시된 방법을 통해 산출된 평가 값은 정비례 관계를 나타내고 있다. 이는 문자 인식율이 높은 영상 문서에 대해서 본 논문에서 제시한 문서 영상 품질 측정 결과 또한 높은 값으로 평가되었음을 알 수 있다.

6. 결론 및 향후 연구

기존 연구에서는 영상 문서의 품질 평가를 위해 문서 내의 가독성을 중점으로 평가한데 비해 본 논문에서는 스캐너를 이용하여 대량의 종이 문서를 스캔하는 과정에서 실제로 발생될 수 있는 기하학적 변형인 수직/수평선 변형, 접힘양 검출, 공백 유무에 따른 회전각을 평가 항목으로 사용하였다. 제안된 문서 영상의 품질을 계산하기 위해서 사용된 항목별 가중치는 300개의 문서 영상을 대상으로 실험한 결과 접힘 변형인 경우 OCR 인식율과 높은 상관 관계를 갖으며 라인 변형의 경우, OCR 인식율과는 가장 낮은 상관 관계를 갖는다.



[그림 8] 각 항목별 평가 값과 인식율 간의 상관 관계

문서의 기울기가 10^0 이상인 경우, 기울기 평가 값에 대해서 0점을 할당하였으나, 이는 본 논문에 실험에 사용된 문자 인식기를 고려한 경우이고, 다양한 한글 인식기를 적용하여 기울기의 측정 기준에 대해서는 원본이 손상되지 않는 범위에서 기울기를 보정 가능한 기울기 값에 대한 보정이 필요하다.

향후 연구로는 문서 자체의 기하학적 변형과 회손 뿐만 아니라 문서 이미지 내 품질을 측정하기 위한 생성 문서의 명료도와 밝기 분포를 측정할 수 있는 방법을 연구 보완하여 보다 다양한 측면에서 생성된 문서 이미지를 평가할 수 있는 방법을 제시하고자 한다.

참고문헌

[1] M. Cannon, P. Kelly, S. Sitharama Iyenger and Nathan Brener, "An automated system for numerically rating document image quality", Proceedings 1997 Symposium on Document Image Understanding Technology, pp. 161-167, 1997.

[2] Henry S. Baird, N, “Document Image Quality : Making Fine Discriminations” Document Analysis and Recognition ICDAR '99, pp. 459-462, 1999.

[3] Souza A, Cheriet M, Naio S, Suen C.Y “Automatic filter Selection Using Image Quality Assesment”, Proceedings of the 7th international conference on document analysis and recognition, pp.508-512, 2003.

[4] Lu, S.J. and Chen, B.M. and Ko, C.C., “Perspective rectification of document images using fuzzy set and morphological operations”, Image and Vision Computing Journal Vol. 23, pp. 541-553, 2005.

[5] Wei Dong, Qian Yu, Zhang C N, Hua Li, “Image Quality Assessment Using Rough Fuzzy Integrals", 27th International conference on Distributed Computing System Workshops, pp. 1-5, 2007.

[6] M. Sezgin, B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation", Journal of Electronic Image, Vol. 13 pp. 146-168, 2004.

[7] 조인성, 이지홍, 오상진, “사용자 중심의 유연한 실시간 머신비전 검사시스템 개발”, 전자공학회논문지, Vol.45 pp. 42-50, 2008.

[8] Hoque, S. Sirlantzis, K. Fairhurst, M.C, “A new chain-code quantization approach enabling high performance handwriting recognition based on multiclassifier schemes”, Document Analysis and Recognition, pp. 834-838, 2003

[9] 이규원, 우동민, “항공영상으로부터 에지 맴의 체인코드 추적에 의한 선소추출”, 한국지능시스템학회논문지, pp. 709~713, 2005.

[10] Nandini N, Srikanta M. K, G. Hemantha, “Estimation of Skew Angle in Binary Document Images Using Hough Transform”, Proceedings of world academy of science, engineering and technology vol 3, pp. 50-55, 2008.

[11] 김진흠, 김민호, “변수선택 편향이 없는 회귀나무를

만들기 위한 알고리즘”, 한국 통계 학회, pp. 459~473, 2004.
[12] 박지선, 김택헌, 류영석, 양성봉 “추천 시스템을 위한 2 way 협동적 필터링 방법을 이용한 예측 알고리즘”, 정보과학회논문지, Vol. 29, pp. 669-675, 2002.
[13] Maya R.Gupta, Nathaniel P, "OCR binarization and image pre-processing for searching historical documents", Pattern Recognition Volume 40, pp. 389-397, 2007.

오 동 열(Dong-Yeol Oh) [정회원]

- 1999년 2월 : 경희대학교 전자계산학과(이학사)
- 2002년 9월 : 숭실대학교 대학원 컴퓨터학과(공학석사)
- 2005년 2월 : 숭실대학교 컴퓨터학과 박사 수료
- 2001년 ~ 현재 : 인젠트 연구개발 본부 차장

<관심분야>
유비쿼터스 컴퓨팅, P2P, 멀티미디어

오 해 석(Hae-Seok Oh) [정회원]

- 1975년 2월 : 서울대학교(공학사)
- 1979년 2월 : 서울대학교(공학석사)
- 1981년 3월 : 서울대학교(공학박사)
- 2000년 2월 ~ 2001년2월 : 미국 스탠퍼드대 학교 객원 교수
- 2004년 2월 ~ 2005년 2월 : 한국 정보처리학회 회장(역임)
- 1982년 2월 ~ 2003년 1월 : 숭실대학교 컴퓨터학부 교수/부총장(역임)
- 2003년 2월 ~ 현재 : 경원대학교 IT 대학 컴퓨터공학과 교수/부총장(역임)

<관심분야>
멀티미디어, 데이터베이스, 지식경영

류 성 열(Sung-Yul Rhew) [정회원]

- 1970년 2월 : 숭실대학교(이학사)
- 1980년 2월 : 연세대학교(공학석사)
- 1996년 8월 : 아주대학교(공학박사)
- 1997년 3월 ~ 1998년 9월 : 조지 맨슨대 객원 교수
- 1998년 3월 ~ 2001년 2월 : 숭실대학교 정보과학대학원 원장
- 1998년 8월 ~ 2004년 7월 : 숭실대학교 전자계산원 원장
- 1999년 3월 ~ 2004년 8월 : 숭실대학교 정보화 지원센터 소장

<관심분야>
소프트웨어 엔지니어링, 시스템 분석 설계, 소프트웨어 보증