필터링 및 피크검출을 이용한 텍스트 추출

*진보라 **조남익

서울대학교 전기·정보공학부

*idealgod@ispl.snu.ac.kr **nicho@snu.ac.kr

Text line extraction based on filtering and peak detection

*Jin, Bora **Cho, Nam-Ik

Department of Electrical and Computer Engineering, Seoul National University

요약

본 논문에서는 문서 영상 처리의 중요한 전처리 과정인 텍스트 라인 추출을 위하여 가우시안 필터링 및 피크 검출을 이용하는 방법을 제안한다. 이는 문서 영상 내의 글자 영역의 픽셀 강도와 텍스트 라인 사이의 간격에 해당하는 강도의 차이로 인해 문서 영상의 각 열마다 높은 피크와 낮은 피크가 번갈아 가며 나타나는 것에 기반으로, 제안하는 알고리즘은 필터 스케일 추정, 필터링 및 피크 검출, 라인 성분 그룹화의 세 단계로 구성된다. 필터 스케일 추정 단계에서는 여러 초기 값으로 필터링하여 피크 차이 간의 히스토그램을 만듦으로써 글자 크기를 대략적으로 예측하며, 필터링 및 피크 검출 단계에서 앞서 예측된 스케일의 가우시안 필터를 이용하여 필터링 한 후, 각각의 열마다 피크를 검출한다. 마지막으로 라인 성분 그룹화를 통하여 검출된 피크를 서로 연결하여 하나의 텍스트 라인을 구성하는 성분들로 그룹화시켜 텍스트 라인을 추출한다. 실험 결과를 통하여, 제안하는 알고리즘은 이진화 과정을 거치지 않음으로써 균일하지 못한 조명환경 등으로 이진화 성능이 좋지 못할 경우에도 텍스트 라인을 추출할 수 있으며, 텍스트 라인 간격이 일정하지 않고 휘어진 라인을 포함하는 경우에도 적용할 수 있음을 확인할 수 있다.

1. 서론

카메라로 취득된 문서 영상에서의 텍스트 라인 추출은 문서의 레이아웃(layout) 분석이나 페이지 분할, 광학식 문자 판독(Optical character Reader) 등의 다양한 문서 영상 처리를 위한 중요한 전처리 과정이다. 특히 문서 영상 처리 기술의 향상을 위해서 보다 정확하고 강인한 텍스트 라인 추출이 요구되고 있기 때문에 지금까지도 라인 추출에 관한 많은 연구가 진행되고 있다. [1]

텍스트 라인 추출 방식에는 대표적으로 투영 프로파일(Projection profile) 기반 방식 [2], 연결 성분(Connected Components)을 이용한 방식이 있다. [3,4] 투영 프로파일 기반 방식은 문서영상의 특정 방향으로 픽셀 값을 누적시켜가며 투영 프로파일을 구함으로써 텍스트 방향을 찾아 라인을 추출하는 기법이다. 이 방식은 비교적 단순하여 널리 사용되어 왔지만, 카메라로 취득된 영상에 대해서는 성능이 떨어질 우려가 있어, 많은 방식들이 주로 텍스트의 연결 성분에 기반 하거나 연결 성분과 투영 프로파일을 함께 이용하고 있다. 연결 성분 기반 방식은 글자들의 연결 요소를 분석하여 라인을 추출하는 기법이다. 이는 텍스트 라인의 형태에 비교적 영향을 덜 받고 성능이 우수하나, 복잡도가 상당히 높은 편이다. 그래서 최근에는 이러한 점들을 보완하고자 문서 영향에 이방성 필터 (anisotropic filter)를 이용하여 텍스트 라인을 추출하는 방법 [5], 특정 필터를 고안하여 라인 추출을 수행하는 방식 [6] 등 필터링에 기반을 둔 방식들이 제안되어 좋은 성능을 보이고 있다.

한편, 기존의 텍스트 라인 추출 방법은 문서 영상을 이진화(binarization)하여 이를 바탕으로 라인 추출을 수행해왔다. 그러나 카메라로 영상 취득 시 그림자가 생기거나 조명의 밝기가 일정하지 않기 때문에 이진화 결과가 나빠지기도 하는데, 이러한 경우 이진화의 성능이 전체 라인 추출 알고리즘 성능에 큰 영향을 미치게 되는 문제점이 있다. 또한 렌즈의 초점이 맞지 않은 영상에 대해서도 라인 추출에 어려움이 있다. 따라서 본 논문에서는 카메라로 취득된 문서 영상에 대해 어떠한 조건에서도 강인한 텍스트 라인 추출 알고리즘을 제안하고자 한다.

제안하는 알고리즘은 문서 영상의 글자 영역의 픽셀 값이 0에 가깝고 여백의 픽셀 값이 흰색에 가깝기 때문에 영역별로 피크(peak)가 나타날 것이라는 점을 이용하고 있다. 즉, 글자 크기에 알맞은 필터링 스케일(scale)을 예측하고 이 스케일로 가우시안 필터링(Gaussian filtering) 시키고 이에 피크 검출을 적용함으로써 텍스트 라인을 찾는다. 이에 따라 제안하는 방식은 크게 필터 스케일 추정, 필터링 및 피크 검출, 라인 성분 그룹화의 세 단계로 구성되는데, 이진화 과정을 거치지 않음으로써, 잘못된 이진화에 따른 성능 저하를 없애고 외부 조건으로부터 강인한 텍스트 라인 추출 결과를 얻을 수 있다. 또한 제안하는 알고리즘은 문서영상의 텍스트 크기가 일정하지 않더라도 적용할 수 있는 장점이 있다.

2. 제안하는 알고리즘

제안하는 알고리즘은 글자의 크기에 부합하는 스케일로 가우시안 흐림 필터를 적용 시켰을 때 영상의 각각의 열에 대하여 텍스트 라인과 텍스트 라인 사이의 간격 간의 밝기 강도 차이로 뚜렷한 낮은 피크와 높은 피크를 형성할 것이라는데 착안하였다. 즉, 문서 영상에 대해

특정 스케일로 가우시안 필터링을 수행한 후 영상의 각각의 열 마다 주변보다 픽셀 강도가 높거나 낮은 피크를 검출하면 텍스트 라인 부분에서 낮은 피크가 나타나고, 라인 간격 부분에서 높은 피크가 거의 일정한 간격으로 번갈아 가며 나타나는 것을 이용하여 텍스트 라인을 검출할 수 있다.

다만, 여기서 중요한 것은 문서 영상에 알맞은 크기의 필터인데, 이를 잘못 찾을 경우 라인 성분을 제대로 찾지 못하므로, 무엇보다 글자의 크기를 잘 반영할 수 있는 필터 스케일을 예측해야 한다는 것이다. 이를 위하여 우선 임의의 값으로 여러 개의 초기 필터링 스케일을 설정하고 이에 따라 가우시안 흐림 필터링 후 각각의 열에 대해 피크 검출을 수행한다. 이때 피크 검출은 주변 값들보다 평균보다 큰 경우 높은 피크로 검출하고 작을 경우 낮은 피크로 검출하는 단순한 방식을 이용한다. 검출된 낮은 피크들은 주로 간격이 균일하여 연속하는 낮은 피크 간의 간격(y 좌표의 차이) 값으로 히스토그램을 생성하면 특정 값에 빈도가 집중되게 되는데, 가장 빈도수가 높은 세 개의 값들의 평균이 텍스트 라인 간격인 것으로 간주하여 스케일로 설정한다.

그리고 보다 다양한 글자 크기의 라인 성분 검출을 위하여 앞서 설정된 스케일뿐만 아니라 이를 상수 배하여 보다 큰 스케일부터 차례로 필터링 시키며 피크를 검출한다. 그러면 각각의 열마다 검출된 낮은 피크가 텍스트 라인의 후보에 해당하는 것으로서 이 지점이 실제 텍스트 라인에 해당할 경우 문서 영상의 열 별로 연속적으로 나타나기 때문에 각각의 피크마다 이웃하는 가장 가까운 낮은 피크와 차례로 연결함으로써 텍스트 라인 성분을 구성할 수 있다.

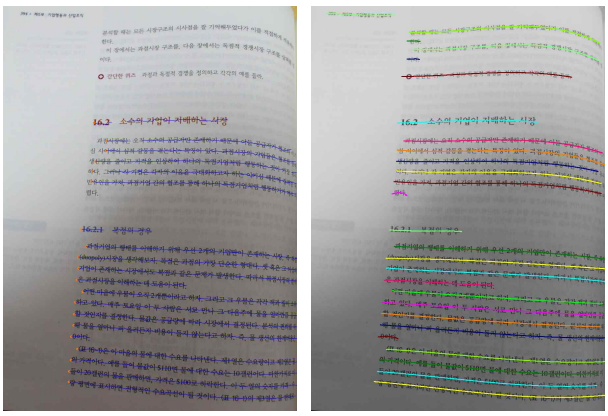
구성된 라인 성분은 텍스트 라인의 일부가 되는 것으로 이들은 하나의 라인을 구성하는 성분끼리 연결될 필요가 있다. 즉 라인 성분에 대하여 가장 가깝고 이웃하는 라인 성분을 찾는 것으로 그룹화될 수 있다. 이때 제한 조건을 두고 가까운 라인 성분을 찾아 연결하는데, 즉 주어진 라인 성분의 오른쪽 영역에서만 연결시킬 라인 성분을 찾게 되며, 주어진 라인 성분을 구성하는 피크 지점 중 제일 오른쪽에 있는 지점의 x좌표가 이웃하는 라인 성분의 가장 왼쪽에 있는 성분의 x좌표보다 크기가 작을 경우에만 연결한다. 또한 라인 성분의 길이가 특정 문턱 값(페이지 너비의 5 %) 보다 짧을 경우에는 텍스트 라인이 아닌 것으로 간주한다.

연결된 라인 성분들을 바탕으로 텍스트 라인을 구성하며, 커브 피팅(curve fitting)을 통하여 최종 텍스트 라인을 얻을 수 있다.

3. 실험결과

제안한 알고리즘을 통해 텍스트 라인을 추출한 결과는 그림 1과 같다. 주어진 문서 영상은 총 26개의 텍스트 라인을 가지고 있는 영상으로 [6]의 알고리즘으로 20개의 라인을 추출하였으며, 제안한 알고리즘의 경우 26개의 라인을 모두 찾는 것을 알 수 있다. 이때 주어진 영상은 문서 내의 밝기가 균일하지 못하고 그림자가 부분마다 존재하였으며 텍스트 라인 간격이 균일하지 않는 부분이 있는 것을 알 수 있는데, 제안하는 알고리즘은 이러한 조건에서도 좋은 결과를 얻음을 확인할 수 있다.

4. 결론



(a) (b)

그림 1 텍스트라인 추출 결과 (a) [6] 결과 (20/26 = 77.0 %) (b) 제안하는 알고리즘 (26/26 = 100 %)

본 논문에서는 글자 크기를 대략적으로 예측하여 이에 알맞은 스케일의 가우시안 필터를 이용하여 필터링 한 후, 피크 검출 및 라인 성분 그룹화를 통하여 텍스트 라인을 추출하는 알고리즘을 제안하였다. 이는 문서 영상 내의 글자 영역의 픽셀 강도와 텍스트 라인 사이의 간격에 해당하는 강도의 차이로 인해 문서 영상의 각 열마다 높은 피크와 낮은 피크가 번갈아 가며 나타나는 것에 기초한 것으로, 실험 결과를 통하여 제안하는 알고리즘은 이진화 단계를 거치지 않고 필터링 함으로써 균일하지 못한 조명환경이나 텍스트 라인 간격이 일정하지 않고 휘어진 라인을 포함하는 경우에도 적용 가능함을 확인할 수 있다.

감사의 글

이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2009-0083495).

참고문헌

[1] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," International Journal on Document Analysis and Recognition, vol. 9, no. 2, pp. 123-138, April 2007.

[2] N. Ouwayed and A. Belaïd, "Multi-oriented text line extraction from handwritten arabic documents," 8th IAPR International Workshop on Document Analysis Systems-DAS'08, 2008.

[3] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri, D.K. Basu, "Text line extraction from multi-skewed handwritten documents," Pattern Recognition, vol. 40, no. 6, pp. 1825-1839, June 2007.

[4] H. I. Koo and N. I. Cho, "State estimation in a document image and its application in text block identification and text line extraction," in European Conference on Computer Vision, ECCV, 2010.

[5] S. S. Bukhari, T. M. Breuel, and F. Shafait, "Textline information extraction from grayscale camera-captured document images," in IEEE International Conference on Image Processing, ICIP, 2009.

[6] S. I. Ha, B. Jin, and N. I. Cho, "Fast text line extraction in document images," in IEEE International Conference on Image Processing, ICIP, 2012.