

저화질 문서영상들을 위한 적응적 이진화 알고리즘

주재현*, 오정수^o

An Adaptive Binarization Algorithm for Degraded Document Images

Jae-hyon Ju*, Jeong-su Oh^o

요약

본 논문은 인쇄체 한글과 한자를 포함하는 저화질 문서영상에 효과적인 이진화 알고리즘을 제안한다. 가는 가로획과 굵은 세로획으로 구성된 문자 속성 때문에 기존 알고리즘은 저화질 문서영상에서 세로획보다 약한 성분을 갖는 가로획을 쉽게 검출할 수 없다. 제안된 알고리즘은 전방향 참조 적응적 알고리즘에 수직 방향 참조 알고리즘을 추가하여 기존 알고리즘이 갖는 문제를 해결한다. 실험 결과는 제안된 알고리즘이 다양한 저화질 문서영상들에서 문자를 잘 추출하는 것을 보여주고 있다.

Key Words : adaptive binarization, document image, characters, degraded image

ABSTRACT

This paper proposes an adaptive binarization algorithm which is highly effective for a degraded document image including printed Hangul and Chinese characters. Because of the attribute of character composed of thin horizontal strokes and thick vertical strokes, the conventional algorithms can't easily extract horizontal strokes which have weaker components than vertical ones in the degraded document image. The proposed algorithm solves the conventional algorithm's problem by adding a vertical-directional reference adaptive binarization algorithm to an omni-directional reference one. The simulation results show the proposed algorithm extracts well characters from various degraded document images

I. 서론

최근 스마트폰 같은 모바일 기기 사용이 대중화됨에 따라 모바일 기기 카메라를 이용한 명함 인식, 문서 인식과 같은 스마트폰용 문자 인식 어플리케이션들이 소개되고 있다. 하지만 모바일 카메라는 일반적으로 화소수는 많지만 카메라 렌즈가 작아서 충분한 광량을 확보하지 못하고 촬영 환경과 촬영 기술로 인해 쉽게 흐려진 저화질 영상이 획득된다. 특히 한글과 한자와 같이 가로획이 가늘고 세로획이 굵은 특징을 갖는 인쇄체 문자들은 포함하는 영

상에서 저화질 영상으로 취득될 때 같은 문자 객체 인데도 불구하고 굵은 세로획에 비해 가는 가로획은 더욱 심하게 영향을 받는다.

한편 문자 인식에서 문서영상의 이진화는 필수적이고 인식율에 직접적으로 영향을 주는 핵심 요소이다. 문서영상의 이진화 알고리즘은 문서 전체에 하나의 임계값을 사용하는 전역적 이진화 알고리즘과 관심화소의 주변 화소들의 정보를 이용해 화소 단위 임계값을 결정하는 지역적 혹은 적응적 이진화 알고리즘으로 구분된다. 전역적 이진화 알고리즘으로 히스토그램의 분산값을 이용해 임계값을 찾는

* 이 논문은 2011학년도 부경대학교의 지원을 받아 수행된 연구임 (PK-2011-66)

• 주저자 : 부경대학교 이미지시스템공학과, jujewang@naver.com, 학생회원

o 교신저자 : 부경대학교 이미지시스템공학과, ojs@pknu.ac.kr, 종신회원

논문번호 : KICS2012-04-211, 접수일자 : 2012년 4월 19일, 최종논문접수일자 : 2012년 6월 14일

Otsu^[1] 알고리즘이 있고, 적응적 이진화 알고리즘으로 주변 화소들의 최대 및 최소 명도를 이용해 임계값을 찾는 Bernsen^[2] 알고리즘, 주변 화소들의 평균과 표준편차를 이용한 Niblack^[3] 알고리즘, Niblack의 단점을 보완한 Sauvola^[4] 알고리즘, 주변 화소들의 평균만 이용하여 고속으로 이진화하는 Bradley^[5] 알고리즘 등이 있고, 최근 처리시간을 고려해 전역과 지역이 혼합된 알고리즘이 소개되고 있다^[6]. 전역적 이진화 알고리즘은 Otsu 알고리즘이, 적응적 이진화 알고리즘은 Sauvola 알고리즘이 가장 우수한 것으로 기술되고 있다^[7,8].

본 연구에서 대상영상은 모바일 기기의 카메라로 획득한 저화질 영상으로 초점이 잘 안 맞거나 모바일 기기의 흔들림으로 인해 흐려지는 현상이 심하게 나타나고, 조명 부족 및 주변 물체에 의한 그림자의 영향으로 전체적으로 명도가 불균일하다. 전체 영상 히스토그램을 구하여 배경과 문자간 분산값을 이용하여 임계값을 정하는 Otsu 알고리즘은 명도 불균일한 영상에서 사용할 수 없고, 전방향 주변 화소들의 정보를 이용해 적응적으로 화소단위의 임계값을 결정하는 적응적 이진화 알고리즘은 명도 불균일에 적응적으로 잘 대응하나 같은 문자 객체에 다른 정도로 흐려지는 저화질 영상에서 강한 세로획의 영향으로 가는 가로획을 소실시키고 있다. 본 논문에서 제안하는 이진화 알고리즘은 기존 적응적 이진화 알고리즘의 문제를 해결하기 위한 것으로 첫 번째 단계에 전방향 주변 화소들을 참조하는 기존 적응적 이진화 알고리즘을 이용해 배경과 분명히 구분되는 세로획 위주의 문자를 검출하고, 두 번째 단계에 강한 세로획의 영향을 배제하기 위해 수직방향 주변 화소들만을 참조하는 적응적 이진화 알고리즘을 수행하여 첫 번째 단계서 손실된 가로획들을 검출한다. 다양한 저화질 문서영상들에서 제안된 알고리즘은 문자들을 적절하게 검출하는 것을 보여주고 있다.

II. 기존 이진화 알고리즘

적응적 이진화 알고리즘 중 가장 성능이 우수한 Sauvola 알고리즘은 관심화소 x 에서 이진화를 위한 임계값 T 를 식(1)와 같이 계산한다.

$$T(x) = m(x) \left[1 + k \left(\frac{s(x)}{R} - 1 \right) \right] \quad (1)$$

여기서 $m(x)$ 과 $s(x)$ 는 각각 관심화소를 중심으로 한 전방향 참조영역인 $w \times w$ 크기의 블록 내 화소들의 평균과 표준편차이고, k 와 R 은 지역 영상의 평균과 표준 편차를 고려하는 상수들로 원 논문에서 각각 0.5와 128를 사용하고 있다.

적응적 이진화 알고리즘의 계산량 문제를 보완한 Bradley 알고리즘은 관심화소에서 이진화를 위한 임계값을 식(2)와 같이 계산한다.

$$T(x) = m(x) - c \quad (2)$$

여기서 $m(x)$ 는 식(1)에서와 같은 의미를 갖고, c 는 상수로 원 논문에서는 5를 사용하고 있다. Bradley 알고리즘은 고속으로 $m(x)$ 를 계산하기 위해 누적합 영상을 이용하고 있다.

III. 문자영상 특성

한글이나 한자의 대부분 문자는 가로획보다 세로획을 더 많이 포함하고 있다. 그래서 폰트를 만들 때 그림 1의 폰트에 보이는 것처럼 가로획은 가늘고 세로획은 굵게 만들어진다.



그림 1. 폰트
Fig. 1. Font

가늘고 굵은 획들로 구성된 문자를 포함한 문서 영상을 취득할 때, 여러 요인으로 저화질 영상이 생성되면 그림 2의 촬영된 영상에 보이는 것처럼 같은 문자 구성 요소인데도 불구하고 굵은 세로획은 흐려지면서도 문자 명도가 유지되지만 가는 가로획은 흐려지면서 문자와 배경의 중간 명도를 갖게 된다. 즉 흐려진 저화질 문서영상에서 문자는 약한 문자성분인 가로획과 강한 문자성분인 세로획으로 구성된다.

저화질 문자영상 특성에 의해 그림 2에 표시된 사각형 영역과 같이 관심화소 주변에 강한 세로획과 약한 가로획이 동시에 포함되는 영역에서 적응적 이진화 알고리즘을 수행하면 약한 세로획의 명도가 강한 가로획보다 배경 명도에 근접해 배경으로 결정될 것이다. 그림 3은 저화질 영상에 Sauvola

알고리즘을 적용해 이진화된 문자영상 예이다. 한글 ‘울’의 경우 ‘ㅓ’의 약한 가로획이 강한 세로획과 동시에 참조영역에 포함되어 배경으로 결정되고 있다. 한자 ‘徒’의 경우 강한 세로획 옆의 약한 가로획은 배경으로 선택되거나 약한 가로획이 강한 세로획에서 멀어지면 참조영역에 강한 세로획이 적게 포함되어 약한 가로획도 정상적으로 문자로 결정된다. 저화질 영상에서 이런 결과는 기존 적응적 이진화 알고리즘들이 공통적으로 갖는 문제점이다.

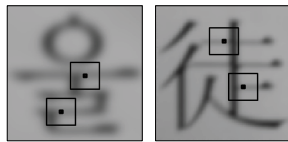


그림 2. 흐려진 문자영상
Fig. 2. Blurred character image



그림 3. Sauvola 알고리즘에 의한 이진화된 문자영상
Fig. 3. Binary character image by Sauvola algorithm

IV. 제안된 적응적 이진화 알고리즘

본 논문에서는 기존 적응적 이진화 알고리즘에서 문자의 가로획이 소실되는 문제점을 보완하는 개선된 적응적 이진화 알고리즘을 제시한다. 그림 4는 제안된 알고리즘의 블록도를 보여주고 있다. 입력영상에서 누적영상을 생성하여 전방향 참조 적응적 이진화 알고리즘을 적용하여 이진영상을 생성하고, 다시 수직방향 참조 적응적 이진화 알고리즘을 수행하여 이진영상을 생성한다. 두 이진영상을 합성하여 최종 이진영상을 완성한다.

4.1. 누적영상 생성

누적영상은 적응적 이진화를 수행하는 과정에서 반복되는 참조영역의 평균 계산을 고속으로 처리를 위해 요구된다. 누적영상 $A(X, Y)$ 은 원영상 $I(\cdot)$ 를 이용해 식(3)와 같이 계산되고, 누적영상을 이용한 관심화소 $I(x, y)$ 에서 참조영역의 평균 $m(x, y)$

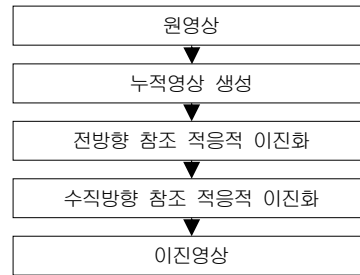


그림 4. 적응적 이진화 블록도
Fig. 4. Block diagram for an adaptive binarization

은 식(4)와 같이 계산된다.

$$A(X, Y) = \sum_{x=0}^X \sum_{y=0}^Y I(x, y), \quad X, Y: 0 \sim N-1 \quad (3)$$

$$m(x, y) = \frac{1}{n} (A(x + \frac{w}{2}, y + \frac{w}{2}) - A(x - \frac{w}{2}, y) - A(x, y + \frac{w}{2}) + A(x - \frac{w}{2}, y - \frac{w}{2})) \quad (4)$$

여기서 w 과 n 은 각각 참조영역의 크기와 화소수이다.

4.2. 전방향 참조 적응적 이진화

전방향 참조 적응적 이진화는 세로획 위주의 분명한 문자를 검출하는 것을 목적으로 하고 추가적인 이진화가 수행되므로 이진화 성능을 높이기 위해 복잡한 알고리즘을 적용하기보다 단순하고 분명한 문자를 검출할 수 있는 알고리즘이 필요하다. Sauvola 알고리즘은 저화질 영상의 이진화에서도 뛰어난 성능을 보이고 있다 그러나 모든 화소에서 임계값 계산을 위해 참조영역의 평균과 분산 계산이 요구되고, 화질에 따라 적절한 파라미터 2개를 결정해 주어야 하는 문제를 갖고 있다. 그래서 본 논문에서는 임계값 결정을 위해 평균만 이용하고 단일 파라미터를 갖는 보다 단순한 Bradley 알고리즘을 전방향 참조 적응적 이진화에 적용한다.

그림 5는 저화질 문서영상에 대하여 Sauvola 알고리즘과 Bradley 알고리즘을 적용해 생성한 이진영상이다. Sauvola 알고리즘에서 파라미터 k 와 R 을 위해 각각 0.05와 128로 사용하였고, Bradley 알고리즘에서 파라미터 c 을 위해 5를 사용하였다. w 는 글자크기의 1/4이 적합하다. 이진영상들은 모두 강한 세로획과 만나는 약한 세로획에서 문자가 손실되는 문제를 보이고 있다. 결과적으로 두 알고리즘은 동등한 이진영상을 보인 반면 Bradley 알고리즘의 계산량은 매우 적다.

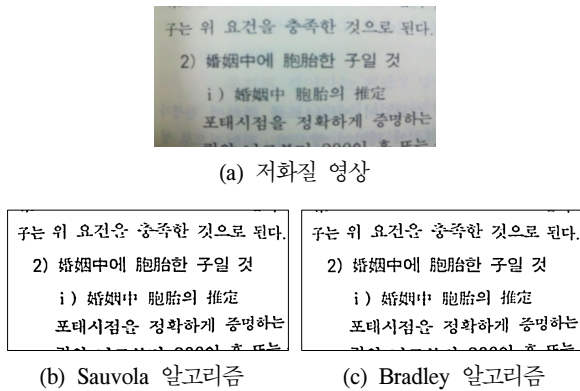


그림 5. 전방향 참조 적응적 이진화.
Fig. 5. Omni-direction reference adaptive binarization.

4.3. 수직방향 참조 적응적 이진화

수직방향 참조 적응적 이진화는 전방향 참조 적응적 이진화에서 강한 세로획 주변의 약한 가로획이 소실되는 단점을 보완하기 위한 추가적인 적응적 이진화이다. 수직방향 참조 적응적 이진화는 Bradley 알고리즘을 그대로 사용하는데 그림 6에 보이는 것처럼 검은 점으로 표시된 관심화소에서 임계값을 결정할 때, 강한 세로획의 영향을 제거하기 위해 참조영역을 사각형으로 표시된 수직 방향으로 제한한다. 강한 세로획이 참조영역에서 제외되어 약한 가로획을 훨씬 민감하게 추출할 수 있다. 본 논문에서 참조영역은 $w \times 1$ 로 w 는 문자크기의 $1/8$ 이 적합하다. 최종 이진영상은 두 번의 이진화에서 생성된 두 이진영상을 합성하고 고립점을 제거하여 완성한다.

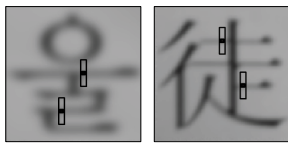


그림 6. 수직방향 참조
Fig. 6. Vertical-directional reference

V. 실험 및 고찰

제안된 알고리즘의 성능 평가를 위해 스마트폰을 이용해 조명 및 흔들림 등이 포함된 다양한 촬영 환경에서 문서영상을 취득하여 제안된 알고리즘을 적용하였다. 제안된 알고리즘은 전방향 참조 적응적 이진화를 위해 참조영역과 상수 c 는 각각 16×16 와 5 가 사용되었고, 수직방향 참조 적응적 이진화를 위해 참조영역과 상수 c 는 각각 8×1 와 4 가 사용되었다. 단 문서영상의 문자 크기가 64×64 정도이

다. 그림 7은 저화질 문서영상에서 제안된 방법을 적용하여 얻은 이진영상을 일부 보여주고 있고, 그림 8은 생성된 이진영상에서 문자 일부가 소실되는 것을 분명히 보여주기 위해 문자단위 영상을 보여주고 있다. 그림 7과 그림 8의 (b)는 전방향 참조 적응적 이진화의 결과로 기존 적응적 이진화 알고리즘과 동등한 결과로 강한 세로획 주변의 약한 가로획이 소실되는 것을 확인할 수 있다. 그림 7과 그림 8의 (c)는 수직방향 참조 적응적 이진화를 통해 가로획을 검출한 결과로 전방향 참조 적응적 이진화에서 소실되었던 약한 가로획을 충실히 검출하는 것을 보여주고 있다. 그림 7과 그림 8의 (d)는 (b)와 (c)의 결과 이진영상을 합성하고 고립점을 제거한 이진영상으로 기존 적응적 이진화 알고리즘의 결과와 동등한 (b) 영상과 비교할 때 가로획을 거의 완전히 추출하는 것을 보여주고 있다. 표 1은 원영상을 저화질 영상으로 변화 후 이진화했을 때 손실되는 문자화소 수를 평가하였다. 저화질 영상은 포토샵에서 '가우시안 블러'의 화소 변수를 조정하여 생성하였다. 제안된 알고리즘은 기존 알고리즘에 비해 15%~33% 정도만 손실되고 있다.

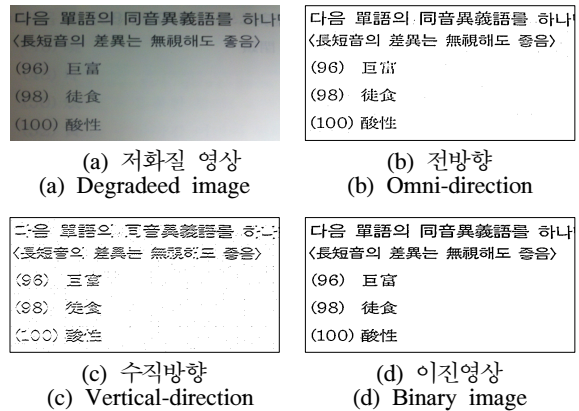


그림 7. 문서영상의 이진화
Fig. 7. Binarization of a document image

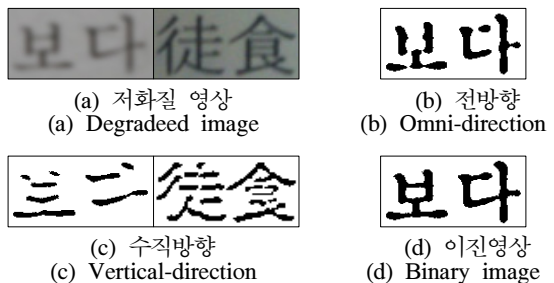


그림 8. 문자영상의 이진화
Fig. 8. Binarization of a character image

표 1. 손실된 문자화소의 수
Table 1. Number of lost character pixels

| 화소 변수 | Sauvola 알고리즘 | 제안된 알고리즘 |
|-------|--------------|----------|
| 1.5 | 3258 | 569 |
| 2.0 | 7236 | 1107 |
| 2.5 | 12224 | 4076 |

VI. 결 론

본 논문에서 저화질 문서영상의 문자인식을 위한 전처리 단계인 이진화 알고리즘을 제안했다. 기존 적응적 이진화 알고리즘들이 저화질 문서영상에서도 우수한 이진영상을 생성하고 있다. 하지만 문자 구성의 특성상 가로획이 가늘고 세로획이 굵은 인쇄체 한글과 한자영상에 기존 적응적 이진화 알고리즘을 적용 시 굵은 세로획 주변의 가는 가로획이 많이 소실되는 문제를 갖는다. 이 문제를 보완하기 위해 본 논문에서는 전방향 참조 적응적 이진화 알고리즘과 수직방향 참조 적응적 이진화 알고리즘과 동등한 알고리즘으로 가로획이 손실된 이진영상을 생성하고 후자는 강한 세로획의 영향을 배제하기 위해 참조영역을 수직방향으로 제한하여 손실된 가로획을 포함한 가로획을 생성한다. 두 알고리즘에 의해 생성된 두 이진영상을 합성하여 최종 이진영상은 생성한다. 다양한 저화질 영상을 이용한 실험로 부터 제안된 알고리즘은 일반 문서영상은 물론 한글과 한자가 포함된 저화질 문서영상에서도 우수한 이진영상을 생성하고 있다.

참 고 문 헌

[1] N. Otsu, "A threshold selection method from grey level histogram," *IEEE Trans. SMC* Vol. 9, No. 12, pp. 62-66, 1979.

[2] J. Bernsen, "Dynamic thresholding of grey-level images," *Proc. ICPR'86*, Vol. 2, pp.1,251-255, Paris, 1986.

[3] W. Niblack, *An Introduction to Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, pp.115-116, 1986.

[4] J. Sauvola, M Pietikainen, "Adaptive document image binarization," *Pattern Recognition*, Vol. 33, No. 2, pp. 225-236, 2000.

[5] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of Graphics Tools*, Vol. 12, No. 2, pp. 13 - 21, 2007.

[6] X. Zheng, W. Tan, and J. Du, "A Fast Adaptive Binarization Method Based on Sub Block OSTU and Improved Sauvola," *Proc. WiCOM*, pp. 1-5, Wuhan, 2011

[7] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, Vol. 13, No. 1, pp. 146 - 165, 2004.

[8] E. Badekas and N. Papamarkos, "Automatic evaluation of document binarization results," *Proc. CIARP'05*, Vol. 3773, pp. 1005 - 1014, Havana, 2005.

주 재 현 (Jae-hyon Ju)



2010년 2월 부경대학교 이미지 시스템공학과 졸업
2010년 3월~현재 부경대학교 대학원 이미지시스템공학과 석사과정
<관심분야> 영상압축, 문서영상처리

오 정 수 (Jeong-su Oh)



1992년 8월 중앙대학교 대학원 전자공학과 석사 졸업
2001년 8월 중앙대학교 첨단 영상대학원 영상공학과 박사
2002년 3월~현재 부경대학교 이미지시스템공학과
<관심분야> 디지털영상처리,

적외선 신호처리