[5-2]

문자와 영상이 혼재된 문서에서의 개선된 이치화 방법

강 구수, 이 수열, 김 서규, 김 재호, 조 석팔

삼성전자 정보통신 연구소

An Improved Binarization Method Of Mixed Text/Image Documents

Gahang Goo Soo, Lee Soo Yeol, Kim Seo Kyu, Kim Jae Ho, Cho Sok Pal

Information Systems Business Div. R & D Center
Samsung Electronics Co.

ABSTRACT

An improved binarization method which seperates text area from image area in mixed text/image documents is proposed and the block artifact of block adaptive thresholding method is removed. A modified laplacian operator is used for classifying text/image area, and continuity also is considered for improving quality of binarized document.

Consequently, the proposed method provides good visual quality in image as well as text area. Moreover, It requires less memory space than the conventional method (at least 4 times).

I. 서 론

원 문서를 흑 백의 이치 정보만을 이용하여 효과적으로 표현하는 방법에 대한 관심이 나날이 증가되어 왔다. 이들은 영상 문서에 대해 공간적인 흑 백의 분포를 이용해 실제 연속적인 밝기를 의사(pseudo) 밝기로 표현하는 디더링(dithering) 방법[1-2]과 문자 문서에 대해 문자의 판독이 용이하도록 문자와 배경을 적절히 분할하는 이치 분할법(bi-level segmentation)에 대한 것[3]으로 크게 분류되어진다. 그러나, 이러한 방법들은 실제적으로 흔히 접하게되는 문자와 영상이 혼재된 문서에 대해서는 그렇게 효율적인 방법이 될 수 없다. 이치 분할법의 적용 경우에는 영상에서 오경계(false contour)등의 문제가 발생하며, 디더링 방법에서는 의사 밝기가 에지(edge)를 몽롱화(smoothing)시켜 문자 부분의 판독성을 저하하는 경향이 있다. 따라서, 근래에 와서는 이러한 경우에서도 효과적인 표현을 위한 연구가 진행되고 있다.

문자와 영상이 혼재된 문서에서는 문자 부분과 영상 부분을 적절히 식별하여 각각의 통계적 특성을 살릴 수 있는 처리가 바람직하다. 이러한 목적으로 Tetsutani와 Ochi는 BAT(block adaptive thresholding) 방법을 제안하였다.[4-5] 이 방법은 처리하고자 하는 원 문서를 먼저 NxN 블록으로 서로 겹치지않게 분리한 뒤, 각각의 블록내부에 속하는 화소(pixel)들이 갖는 밝기값(gray level)의 최대값과 최소값의 차만으로 그 블록이 문자 부분에 속하는지 영상 부분에 속하는지를 식별하여 문자 부분일 때는 블록내의 화소들에 이치 분할법을 적용하고, 영상 부분일 때는 화소들에 디더링 방법을 적용함으로써 이전의 방법들보다 효과적인 표현이 가능하도록 하였다. 그러나, 이 방법은 국부적인 최대 밝기값과 최소 밝기값을 이용하므로 비교적 간단하지만, 블록단위의 처리로 인하여 서로 다른 부분으로 식별된 블록과 블록의 경계에서 시각적으로 눈에 거슬리는 블록화 현상이 나타나는 문제가 있다. 뿐만아니라 처리할려는 문서의 어떤 블록이 문자를 포함하지 않는 배경 또는 블록 크기보다 두꺼운 폭을 갖는 문자 내부에 설정된다면, 이를 영상 부분으로 식별하여 블록내부의 화소들에 대해 디더링을 하게된다. 이때 흑 또는 백점이 발생될 수 있으며, 이러한 결과는 흑 백의 런장(run-length)을 짧게 하여 런장 부호화시 부호화 효율을 감소시킨다.

본 논문에서는 하드웨어(hardware) 구현이 용이하면서도 BAT에서 발생되는 문제점들을 효율적으로 억제하는 화소 단위로 처리하는 방법을 제안하였다. 이 방법은 현재 처리하고자 하는 화소에 대해 MxL의 국부적인 창(window)을 설정하여 그 창내의 화소들이 갖는 최대 밝기값과 최소 밝기값을 이용하여 현재 화소가 문자 부분에 있는지 영상 부분에 있는지를 식별한다. 또한, 문자를 창내에 포함하지 않는 배경과 창보다 큰 폭을 갖는 문자 내부의 화소들을 효과적으로 처리하기 위하여 현재 화소가 배경 부분에 있는지와 문자 내부 부분에 있는지도 식별하여 그 특성에 맞게 처리를 한다. 그리고, 영역 식별의 정확도를 보다 향상시키기 위하여 현재 화소가 영상 부분에 존재하는 것으로 식별될 때에는 영상 영역에 속하는 화소는 독립적으로 존재할 수 없다는 전제아래 인접한 화소들과 동일한 특성을 갖는지 연속성을 조사하였다.

II. 종래의 간이 식별 방법

일반적으로 문서에서 문자 부분은 영상 부분과는 달리 문자와 배경의 경계에서 큰 밝기 변화를 가진다는 사실에 기반을 둔 종래의 방법에 대한 알고리듬은 아래와 같다.

단계 1. 처리하고자 하는 원 문서를 NxN 화소들로 구성되는 블록들로 서로 겹치지 않게 분리한다.

단계 2. 각 블록에 대해 그 블록에 속하는 화소들의 밝기값들 중 최대값과 최소값을 구한다.

단계 3. 블록내의 최대 밝기값과 최소 밝기값의 차가 미리 정의된 역치 P보다 크면 이 블록을 문자 부분이라고 식별한다. 따라서, 이 블록내의 각 화소에 대해 이치 분할법을 적용한다.

단계 4. 블록내의 최대 밝기값과 최소 밝기값의 차가 P보다 작으면 이 블록을 영상 부분이라고 식별한다. 따라서, 이 블록내의 각 화소들에 대해 디더링 방법을 적용한다.

상술된 알고리듬은 단계 3, 4와 같이 블록 단위로 각 화소들을 식별하므로 그림 1의 예에서 볼 수 있듯이 서로 다른 부분으로 식별된 인접한 블록간의 경계에서 블록화 현상이 나타나는 경향이 있다. 또한, 블록 K3, K8과 같이 한 블록이 문자 부분에 속하는 특성을 가진 화소들과 영상 부분에 속하는 특성을 가진 화소들로 구성될 때, 블록내의 화소들을 각각의 특성에 맞는 부분으로 구별할 수

없으므로 적절히 처리할 수 없다. 그리고, 블록내에 역치 P 보다 큰 밝기값의 변화가 없다면, 그 블록을 영상 부분으로 식별하여 블록내부 화소들을 디더링한다. 따라서, 보통의 문서에서 흔히 발생하는 경우로 블록 K14와 같이 블록이 문자를 포함하지 않는 배경 부분에 설정될 때, 블록내의 화소가 갖는 밝기 값이 디더 행렬의 가장 높은 역치값보다 낮다면 그 블록내의 화소들을 디더링한 결과 영상은 배경에서 흑점을 갖는다. 또한, 어떤 블록이 블록의 크기보다 큰 폭을 갖는 문자의 내부에 설정될 때, 블록내의 화소가 갖는 밝기 값이 디더 행렬의 가장 낮은 역치값보다 높다면 앞과 마찬가지로 그 블록내의 화소들을 디더링한 결과 영상은 문자 내부에서 백점을 갖는다. 따라서, 이들은 문자와 배경에서의 명확한 대비를 감퇴시켜 결과적으로 시각적인 측면에서의 화질을 손상시킨다. 또한, 이치 데이타의 부호화를 위하여 일반적으로 널리 사용되고 있는 런장 부호화시, 이들의 발생을 효과적으로 방지한 경우에 비하여 전체적인 문서의 평균 백런장(white run-length)과 흑런장의 길이가 작아지므로 높은 압축률을 저해하는 요인이된다.

III. 제안한 간이 식별 방법

본 방법은 종래의 간이 식별 방법에서 발생하는 블록화 현상을 방지하기 위하여 블록 단위의 식별이 아니라 화소 단위의 식별을 한다. 그리고, 종래의 방법보다 효과적으로 각 화소들을 처리하기 위하여 현재 화소가 어떤 부분에 속하는 지를 보다 세밀히 하여 문자 부분, 영상 부분, 배경 부분, 그리고 문자 내부 부분으로 식별한다. 또한, 영상 부분은 공간적으로 인접 화소들간에 연속성을 가진다는 전제아래 일단 영상 부분으로 간주되는 화소에 대해서는 이 화소와 인접한 화소들의 특성이 동일할 때만 실(real) 영상 영역으로 식별한다. 이러한 알고리듬은 아래와 같다.

단계 1. 처리하고자 하는 원 문서의 좌측 상단부터 우측 하단까지 순서대로 한 화소씩 주사한다. 그리고, 현재 주사하는 화소를 현재 화소라 한다.

단계 2. 현재 화소에 대해 MxL의 국부적인 창을 씌운다.

단계 3. 창내의 화소들이 갖는 밝기값들 중에서 최대값과 최소값을 구한다.

단계 4. 앞서 얻은 최소값이 미리 정의된 역치 Tmax 보다 크면, 현재 화소가 배경 부분에 속한다고 식별하여 그 화소를 백으로 처리한다. 다음, 단계 1로 간다.

단계 5. 앞서 얻은 최대값이 미리 정의된 역치 Tmin 보다 적으면, 현재 화소가 문자 내부에 존재한다고 식별하여 그 화소를 흑으로 한다. 다음, 단계 1로 간다.

단계 6. 앞서 얻은 최대값과 최소값의 차이가 어떤 미리 정의된 역치 P보다 크면, 현재 화소가 문자 영역에 속한다고 식별하고 그 화소를 이치 분할법으로 처리한다. 다음, 단계 1로 간다.

단계 7. 앞선 단계들의 각 조건을 만족하지 않는다면 현재 화소를 일단 영상 영역으로 간주하고, 다음 처리될 화소들을 위하여 이 화소의 연속성 상태비트를 1로만든 후, 연속성을 조사한다. 여기서, 연속성은 그림 2의 연속성 고려 마스크에서 X가 현재 화소라 할 때 이미 처리된 인접한 네화소인 A, B, C 및 D의 연속성 상태 비트가 모두 1일 때만 존재하는 것으로 판단한다. 연속성을 가지면 현재 화소가 실 영상 영역에 속한다고 판단하여 디더링법으로 처리하고, 그렇지않으면 이 화소의 밝기값과 미리 정의된 역치 T를 직접 비교하여 그 출력을 결정한다. 다음, 단계 1로 간다.

IV. 실험결과 및 검토

본 논문에서는 문자와 영상이 혼재된 문서에 대해 제안한 방법의 효율성을 검토하고 기존의 방법과의 성능을 비교하고자, IBM/PC-386 상에서 C-언어로 실험을 수행하였다. 이때, 실험 문서로는 실제 스캐너(scanner)를 통하여 수평 수직 방향으로 각각 밀리미터당 7.7 화소로 표본화(sampling)되고, 8 비트로 양자화(quantization)된 512x512 화소 크기의 문서를 사용하였다.

실험은 각 방법으로 처리된 문서들에 대한 주관적인 화질 평가와 실제 부호화시의 최저 한계가되는 엔트로피(entrophy)를 계산하여 비교하였다.

본 결과들에서 각 식별 방법을 통해 영상 부분으로 식별된 화소들에 대해 적용한 디더법은 일반적으로 널리 사용되고 있는 조직적 디더법으로 그림 3과 같은 4x4 BAYER 디더행렬을 사용함으로써 16단계의 의사 밝기를 표현한다. 그리고, 문자 부분으로 식별된 화소들에 대해 적용한 이치 분할법은, 기존의 방법에서 사용했던 것처럼 국부적인 밝기 분포 특성을 무시하고 전 문서에 대해 단일 역치값을 적용하는 방법 대신에, 아래와 같이 국부적인 밝기 특성에 따라 역치값이 자동적으로 변하는 국부 적응 역치화(local adaptive thresholding)를 사용함으로써 복잡도의 증가없이 문자와 배경의 분리를 보다 세밀하고 정확하게 하였다.

IF I(x,y) >= (LOCAL_MAX+LOCAL_MIN)/2, O(x,y)=255;(백)
ELSE O(x,y)=0;(흑) -----(1)

여기서, I(x,y)는 처리하고자 하는 문서에 있어서 x행 y열에 위치하는 화소가 갖는 밝기값(0<=I(x,y)<= 255)이며, O(x,y)는 처리된 문서의 밝기값(0 또는 255)이다. 그리고, LOCAL_MAX 및 LOCAL_MIN은 x행 y열 위치의 화소가 포함되는 블록 또는 그 화소가 중심이 되는 창내의 화소들이 갖는 밝기값들의 최대 및 최소값이다.

실험 영상에 대하여 각 식별 방법을 적용한 결과는 그림 4의 (c)와 (d)에 각각 주어져 있다. 여기서, 기존의 방법에서의 블록의 크기는 블록화 현상과 문자의 선명화를 고려하여 8x8를 사용하였고, 제안한 방법에서의 창의 크기는 하드웨어 구성시 소요 메모리의 감소와 계산량의 감소를 위하여 각 화소의 식별시 그림 2와 같은 창을 사용하였지만, 이 결과는 보다 확장된 창을 이용한 경우에 비하여 거의 유사하였다.

그림 4는 여러 밝기의 배경에서 문자들을 갖는 문자 문서에 대해 처리한 결과들을 제시하며, (a)와 (b)의 결과는 문자와 영상의 식별 처리에 대한 제안한 방법의 효율성을 나타내기 위하여, 단일 역치값을 갖는 이치 분할법과 위에서 설명한 BAYER 디더 행렬로써 각각 처리한 것이다. 그림 4에서 볼 수 있듯이 기존의 식별 방법을 이용하여 처리한 결과인 (c)는, 문자는 어느 정도 (a)의 결과에 유사한 판독성을 가진다. 그러나, 약간 어두운 배경에서 두드러진 블록화와 밝은 배경과 큰 문자 내부에서 반전된 화질에 도움이 되지않는 점들을 갖는다. 이에 대하여, 제안한 식별 방법으로 처리한 결과인 (d)는 (c)에서 보여지는 문제점들이 거의 나타나지 않음을 시각적으로 확인할 수 있다. 그리고, 그림 5는 인물 영상과 지도가 혼합된 문서이며, (a)와 (b)에 제시된 결과는 역시 단일 역치값을 갖는 이치 분할법과 BAYER 디더 행렬로써 각각 처리한 것이다. 이 결과에서 기존의 방법에 의한 결과인 (c)는, 시각적으로 눈에 거슬리는 인물 영상의 여러 곳에 나타나는 블록화와 배경에서 반전된 흑점들의 존재, 그리고 (a)에 비해 떨어지는 문자의 선명화를 갖는다. 이에 대하여, 제안한 식별 방법으로 처리한 결과인 (d)는 앞의 결과와 마찬가지로 우수한 화질을 가짐을 확인할 수 있다.

다음으로 각 처리 방법이 부호화시의 압축률에 어떠한 영향을 미치는 가를 조사하기 위하여 이치 데이타의 부호화법으로 널리 사용되고 있는 런장 부호화시의 이론적인 최고 압축률을 제공하는 평균 엔트로피 H를 그림 4, 5의 결과들에 대해 계산하였다. 이 계산식은 문서내의 흑런장

과 백런장이 서로 독립된 특성을 가진다는 가정 아래 식 (2)로 표현되며, 계산 결과는 표 1에 제시되었다. 이 결과에서 볼 수 있듯이 제안한 방법의 경우는 배경과 문자 내부에서 짧은 런장의 발생을 비교적 작게 하여 평균 흑백런장의 길이(RB와 RW)를 증가시키고, 엔트로피(H)에서도 약 18 퍼센트 감소되었다.

$$H = [N_b \sum_i P_{bi} * \log(1/P_{bi}) + N_w \sum_j P_{wj} * \log(1/P_{wj})]/N_t \quad --(2)$$

$$RB = \sum_i i * P_{bi}, \qquad RW = \sum_j j * P_{wj}$$

여기서, P_{bi}와 P_{wi}는 문서내에서의 흑연속길이와 백연속길이의 발생 확률이며, N_t, N_b 및 N_w는 각각 문서내의 총 화소수, 총 흑점수, 그리고 총 백점수이다.

V. 결 론

본 논문에서는 문자와 영상이 혼재된 문서에 대해 문자 부분과 영상 부분을 간이 식별하는 방법으로써 Tetsutani와 Ochi가 제안했던 BAT방법에서 발생하는 문제점들을 분석하고, 이들이 제거된 개선된 식별 방법을 제안하였다.

제안된 방법은 화소 단위의 간이 식별을 통하여 블록화 현상의 발생을 방지하였을 뿐만아니라 식별시 배경 부분과 문자 내부 부분의 고려와 연속성의 조사, 그리고 효율적인 국부 적응 역치화를 통하여 비교적 시각적으로 우수한 결과를 얻었다. 또한, 실제 런장 부호화시의 이론적인 최저 한계가 되는 엔트로피의 비교로 제안한 방법이 기존의 방법에 비해 효율적인 부호화가 가능함을 확인하였다.

또한, 제안된 방법은 이러한 우수한 성능을 얻으면서도 기존의 방법에 비해 두개의 라인(line) 메모리(8비트와 1비트)만을 가지므로 비용적으로 하드웨어 구현이 훨씬 용이한 이점을 가진다.

참고 문헌

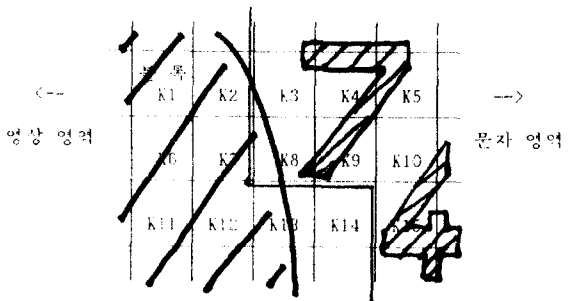
[1] Sahoo et al., " A survey of thresholding techniques," CVGIP 41, 233-260, 1988.

[2] B. E. Bayer, "An optimum method for two-level rendition of continuous-tone pictures," in ICC Conf. Rec., 26-11, 1973.

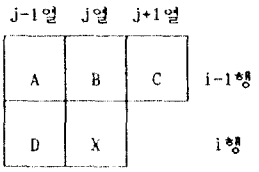
[3] R. A. Ulichney, "Dithering with blue noise," IEEE Proc., VOL. 76, No.1, Jan. 1988.

[4] N. Tetsutani and H. Ochi, "A quasi-tone reproduction method without reducing character image quality," IECE, Japan, Image Eng. Group, IE81-57, Sept. 1981.

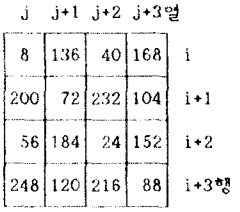
[5] H. Ochi and N. Tetsutani, "A new half-tone reproduction and transmission method using standard black and white facsimile code," IEEE Trans. on comm., VOL. COM-35, No. 4, Apr. 1987.



그 림 1. 종래의 방법의 문제점을 도시하는 예.



그 림 2. 연속성 고려 마스크.



그 림 3. BAYER 4x4 디더 행렬.
(각 좌표에 있는 숫자는 역치임)

표 1. 종래와 제안한 방법을 적용한 그림 4, 5의 결과에서 얻은 평균 흑백 런장의 길이와 평균 엔트로피.

	종래의 방법		제안한 방법	
	4의 (c)	5의 (c)	4의 (d)	5의 (d)
RB	1.30	2.73	1.55	3.37
RW	2.87	2.93	3.30	4.30
H	0.57	0.64	0.48	0.53

(a) 간단한 이치 분할법

(b) 간단한 디더법

(c) 종래의 방법(BAT 이용)

(d) 제안한 방법

그 림 4. 각 방법을 적용한 결과들.

(a) 간단한 이치 분할법

(b) 간단한 디더법

(c) 종래의 방법(BAT 이용)

(d) 제안한 방법

그 림 5. 각 방법을 적용한 결과들.