쿼드트리로 구성된 한글 문서 영상에서의 문자추출에 관한 연구

백 은 경°, 조 동 섭
이화여자대학교 전자계산학과 대학원

EXTRACTION OF CHARACTERS FROM THE QUADTREE ENCODED DOCUMENT IMAGE OF HANGUL

Eunkyoung Paik°, Dongsub Cho
Department of Computer Science, Ewha University

ABSTRACT

In this paper the method of representing the document image by the quadtree data structure, and extracting each character seperately from the constructed quadtree are described. The document image is represented by a binary encoded quadtree and the segmentation is performed according to the information of each leaf node of the quadtree. Then, each character is extracted by the relation of positions of segments. This method enables to extract characters without examining every pixel in the image and the required storage of document image is decreased.

I. 서 론

컴퓨터가 다양한 분야에서 복잡하고 방대한 정보를 처리하게 되면서 키보드에 의한 입력만으로는 처리 능력의 한계를 가져 왔다. 따라서 스캐너나 카메라를 이용하여 정보를 입력할 수 있도록 하는 문서 자동 인식 시스템에 대한 연구가 활발히 진행 중에 있다. 문서 자동 인식 시스템의 개발을 위해서는 문서에서 문자를 추출하는 연구, 추출된 문자를 인식하는 연구, 그리고 인식된 문자를 데이타 베이스화 하는 연구가 요구된다[1]. 그러나 그동안 문자의 인식에만 연구가 집중되어 왔으므로 전처리 분야에 보다 집중적인 연구가 요구되는 실정이다[1].

문서 영상을 처리하는데 있어서 요구되는 여러 가지 사항 중 기억 장소의 양 또한 문제가 된다. 영상 데이타는 높은 해상도를 요구하므로, 이를 저장하기 위하여 많은 양의 기억 장소가 필요하다. 이러한 기억 장소 문제를 해결하기 위하여 쿼드트리(quadtree)를 이용하여 문서 영상을 저장하고 이로부터의 문자 추출을 시도하였다. 쿼드트리는 반복적인 분할(recursive decomposition)을 기본으로 하는 계층적 데이타 구조(hierarchical data structure)로서, 주어진 2차원 입력 영상을 모두 1(흑화소) 또는 0(백화소)만을 갖는 영역이 될 때까지 같은 크기의 4분영상(quadrant)으로 반복하여 나누어 저장한다.

본 연구에서는 문자 추출을 위하여, 2진 코드화된 쿼드트리에 의하여 문서 영상을 저장하고 이로부터 트리의 각 노드의 이웃노드(neighbor)를 찾아서 영상을 분할한다. 분할된 각 절편의 영상 내에서의 위치와 크기에 의하여 개별 문자를 추출하며, 문자 추출 과정에서 섬형 잡상을 자동으로 제거하도록 한다. 실지로 쿼드트리 표현에 의하여 영상 데이타의 기억 장소 요구량이 현저히 감소하였으며, 절편 좌표에 의한 문자 추출은 문서 영상 내의 모든 화소를 조사하는 이전의 방법에 비하여 자료 처리량을 월씬 감소 시켰다.

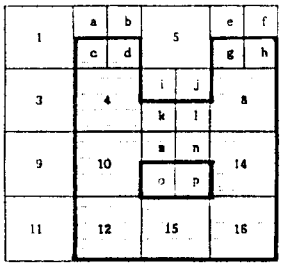
본 연구에서 처리 대상이 되는 문서 영상은 동일한 크기의 인쇄체 문자만을 포함하는 것으로 하였다. 문서는 카메라에 의하여 입력하였으며, 256×256 화상을 기본으로 처리 하였다.

II. 쿼드트리에 의한 영상표현

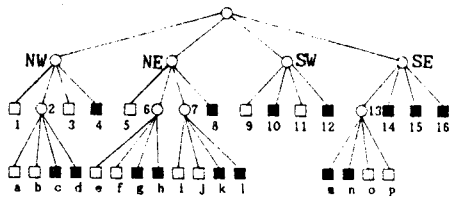
1. 쿼드트리의 구축(Construction)

쿼드트리의 구축(construction)은 크기 $2^k \times 2^k (k \geq 1)$인 2진 영상에서 영역(region)을 얻을 때까지 4분을 반복하면서 진출 차수(out degree)가 4인 트리를 만들어 가는 것이다. 영역은 모두 1 또는 모두 0만을 포함하는 부분 영상을 말하며, 마지막으로 얻어진 영역들은 트리의 단말 노드(leaf node)에 저장한다. 쿼드 트리의 루트 노드(root node)는 전체 영상에 대응되며, 노드의 각 자식 노드(son)는 NW(북서 방면), NE(북동 방면), SW(남서 방면), SE(남동 방면)의 순서로 명명(labeling)된 4분영상(quadrant)을 표현한다. 단말 노드는 그것이 대응하는 4분영상이 완전히 문자 안에 있는지, 완전히 문자 밖에 있는지에 따라 BLACK(검은 영역)과 WHITE(흰 영역)로 구분된다. 단말 노드가 아닌

모든 노드는 GRAY(검은 영역과 흰 영역이 혼합된 영상)로 구분되며, 부분 영상을 나타낸다. 그림 1(a)는 영상을 4분한 예이며 4분된 영상을 쿼드트리로 표현하면 그림 1(b)와 같이 된다.



(a) 영역을 얻기 위해 영상을 4분



(b) 그림 1(a)에 대응하는 쿼드트리

그림 1. 영상을 쿼드트리로 표현하는 예

2. 컴퓨터에 쿼드트리를 표현하는 방법

본 연구에서는 쿼드트리의 저장 기법으로 DF-expression(Depth-First picture expression)을 사용하였다. DF-expression은 쿼드트리의 각 노드를 전위 순회(preorder traversal)함에 의하여 노드의 코드를 얻는 코드화 방식이다. 루트로부터 전위 순회를 시작하여 GRAY 노드를 (, BLACK 노드를 1, WHITE 노드를 0으로 표현하여 보면 그림 1의 문자 영상은 다음과 같이 표현된다.

((0 (0 0 1 1 0 1 (0 (0 0 1 1 (0 0 1 1 1 (0 1 0 1 ((1 1 0 0 1 1 1

노드의 종류가 GRAY, BLACK, WHITE의 3 종류이므로, DF-expression에 의하여 각 노드는 2 비트(bit)에 표현할 수 있다. 그러므로 전체 영상을 저장하기 위해서는 (노드의 수)×2 비트의 기억 장소만 있으면 된다. 노드의 나열 순서가 트리를 전위 순회한 순서이며, 자식 노드들의 순서가 NW, NE, SW, SE라는 것을 알고 있으므로, 이에 의하여 각 노드의 부모 노드와 자식 노드들을 찾을 수 있다.

Ⅲ. 문서 영상 분할

1. 4분 영상의 인접

문자 추출을 위한 전처리로 영상 분할(segmentation)을 수행한다. 분할을 위하여, 트리를 전위 순회하면서 만나게 되는 BLACK 노드들마다 인접한 4분영상을 조사한다. 인접은 다음과 같이 정의되며, 중복된 조사를 피하기 위하여 E(동) 방면과 S(남) 방면에 대해서만 조사한다.

(가) GSN(P,D) = Q 노드 Q는 노드 P의 D 방면에 인접하면서, P에 대응하는 블럭보다 크기가 크거나 같은 블럭 중 가장 작은 블럭(GRAY일 수도 있다)에 대응한다.

(나) CSN(P,D,C) = Q 노드 Q는 노드 P의 C 모서리의 D 방면에 인접한 가장 작은 블럭에 해당한다.

예를 들어, 그림 1에서 GSN(4,E) = 7, GSN(4,S) = 10, CSN(4,E,SE) = k이다.

2. 인접 조사에 의한 영상 분할

영상의 분할은 3단계로 이루어지며, 분할을 위하여 모두 두 번의 트리 순회를 하게 된다. 제 1단계에서는, 가능한 모든 BLACK 노드들의 쌍을 조사하여 BLACK 노드들을 명명(labeling)한다. 명명은 분류 번호를 할당함으로써 이루어지며, 같은 분류 번호를 갖는 성분들의 좌표값을 서로 비교하여 최소값과 최대값을 얻는다. 전체 영상의 시작점과 크기를 알고 있으므로 단말 노드의 상대적 위치에 의하여 그 좌표를 알 수 있다. 이 때, 이미 분류된 영역들 사이의 동치(equivalence)들이 모두 발견되고 이들의 분류 번호들이 동치 쌍(equivalence pair)의 목록(list)에 추가된다.

제 2단계에서는, 1단계에서 만든 동치 쌍의 목록에 의하여 동치 부류(equivalence class)를 생성하고, 생성된 각 동치 부류를 포함하는 최소의 윈도우를 잡는다. 윈도우는 최소, 최대 좌표값에 의하여 정의된다.

제 3단계에서는 트리를 한 번 더 순회하면서 한 동치 부류 내의 연결 요소들에 대하여 같은 분류 번호를 할당한다.

제 1단계를 거친 영상에 적용한 결과를 그림 2에 보였다. 그림 2의 예에 제 2단계를 적용하여 생성한 동치부류의 윈도우는 그림 3과 같으며, 이로부터 제 3단계를 적용하여 분류 번호를 갱신한 결과는 그림 4와 같다.

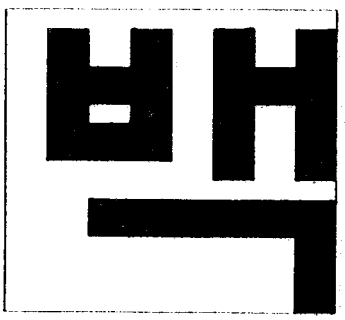


그림 2. (a) 원본 영상

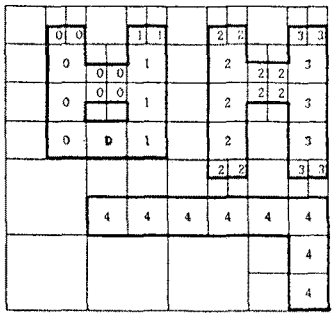


그림 2. (b) 연결 요소들의 분류 번호(제 1단계 적용 결과)

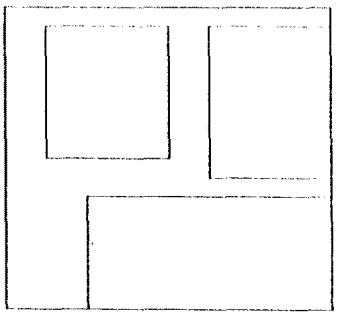


그림 3. 그림 2 영상의 각 절편의 윈도우

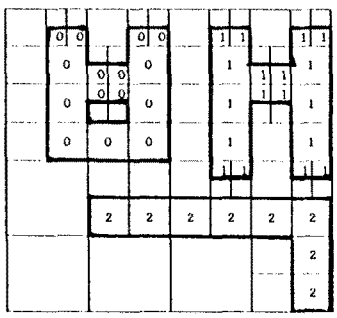


그림 4. 그림 2로부터 분류 번호를 갱신한 결과

3. 잡상의 제거

잡상(noise)이란 문자 인식에 있어 입력 대상의 모양이 왜곡되거나 대상의 크기 또는 위치 변화 등을 의미한다. 잡상에는 돌출형 잡상, 섬형 잡상, 구멍형 잡상, 패인형 잡상의 4가지가 있다[5]. 그림 5에 잡상의 종류를 보였다. 4가지의 잡상 중 돌출형 잡상과 구멍형 잡상, 패인형 잡상은 문자의 절편 내에 포함되는 잡상인 반면에 섬형 잡상은 독립된 절편을 구성한다. 그러므로 문자 인식 과정에서 다른 3가지의 잡상 문제는 자모(alphabet)의 형태 왜곡을 처리하는 과정에서 해결 될 수 있으나, 섬형 잡상에 대해서는 별도의 처리가 요구된다.

영상을 분할하는 과정에서 절편과 그 절편의 윈도우가 구해지므로, 절편의 윈도우의 크기를 조사하면(영상 분할

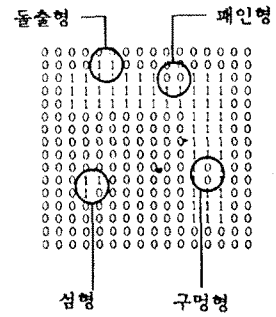


그림 5. 잡상의 종류

제 3단계) 그 절편이 섬형 잡상인지 문자의 절편인지를 판단할 수 있다. 따라서 경계값(threshold) 이하의 크기를 갖는 윈도우의 절편을 잡상으로 판단하고 이를 제거함으로써 섬형 잡상을 제거한다.

Ⅳ. 쿼드트리로 표현된 문서 영상에서 문자의 추출

문자의 추출은 영상 분할 단계에서 얻은 절편의 윈도우를 확장함으로써 이루어진다. 한글의 모든 문자는 C(자음) + V(모음), C + V + C, C + V + C + C, C + V + V, C + V + V + C, C + V + V + C + C의 6가지 형식 중 한 가지로 모아쓰에 의하여 한 문자를 구성한다. 그러므로 모든 문자가 일정한 크기의 윈도우 안에 잡힌다. 문자 추출을 위한 윈도우는 이동할 수 있어야 하므로, 절편의 윈도우와는 달리 가로, 세로의 길이에 의하여 정의된다. 입력되는 문서 영상에서 문자의 크기가 다양하므로, 각 문서에 대하여 동적(dynamic)으로 문자를 위한 윈도우의 크기를 결정해야 한다. 이 때 한 문서 내에 있는 모든 문자의 크기는 동일한 것으로, 가로 쓰기되어 있는 것으로 한다.

영상 분할 단계에서 각 절편들의 윈도우를 최소, 최대 좌표값으로 정의하였으므로 이에 의하여 각 절편의 x축 길이와 y축 길이를 알 수 있다. 그러므로, 절편의 윈도우 값 중 y축 방향 길이의 최대 값과 x축 방향 길이의 최대값을 구하여 문자 추출을 위한 윈도우의 가로, 세로 길이를 정하면 된다.

윈도우의 크기를 정한 후엔 한 윈도우 안에 들어오는 절편들이 모여서 한 문자를 이루는 것으로 결정한다. 그런데, 절편들의 나열 순서는 쿼드트리의 순회 순서이므로 윈도우로 이들을 효율적으로 추출하기 위해서는 전처리로서 정렬(sorting)이 요구된다. 우선 y축에 대하여 정렬함으로써 행들을 각각 구분한다. 행과 행이 구분되면 한 행 안에 있는 절편들을 x축에 대하여 정렬한 후, x축에 대하여 정렬한 절편들에 대하여 윈도우에 의한 문자 추출을 수행한다.

Ⅴ. 구현 및 결과

구현을 위한 한글 문서 영상은 카메라로 입력하였다.

카메라에 아날로그 신호(analog signal)로 입력되는 영상 신호를 디지탈 신호(digital signal)로 변환하기 위하여 SeeEye-256 화상 처리 보드를 IBM-PC/ AT에 장착함으로써, 카메라로 입력한 텔레비젼 영상 신호를 256 명암 단계(gray level)로 디지탈화하여 영상 메모리(memory)에 저장하고, 영상 메모리에 저장된 영상 데이타를 아날로그 값으로 변환하여 256×256의 해상도를 갖는 텔레비젼 모니터(monitor)에 출력하도록 하였다.

문서 영상을 입력받을 때는 흑백(흑은 1, 백은 0)의 상태로만 입력받게 된다[1]. 그러므로, 256 명암 단계로 디지탈화한 영상 자료에 경계값(threshold)을 주어 경계값보다 큰 값을 갖는 화소는 0으로, 보다 작은 값을 갖는 화소는 1로 변환한 영상 데이타를 구현의 대상으로 하였다. 그림 6은 경계값을 50으로 하여 얻은 흑백 영상이며, 256×256 화소들로 표현되는 이 영상을 쿼드트리로 표현하면 9321개의 노드로 구성된다. 이는 그림 7과 같이 분할되며, 이 때 심형 잡상이 제거된 것을 볼 수 있다.

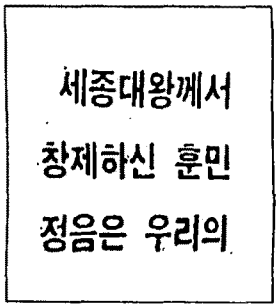


그림 6. 입력한 영상

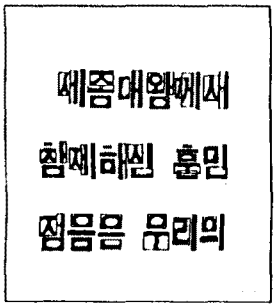
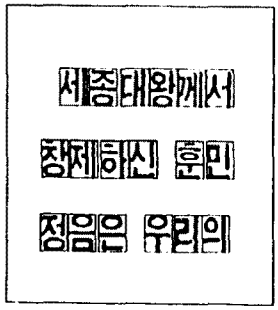


그림 7. 문서 영상의 분할

그림 7의 절편들의 윈도우로부터 그림 8(a)의 문자 추출 윈도우를 얻어서 문자를 추출하면 그림 8(b)와 같이 문맥상의 순서로 개별 문자를 추출할 수 있다.



(a) 문자 추출을 위한 윈도우 (b) 추출된 개별 문자들

그림 8. 개별 문자의 추출

Ⅵ. 결론

중요한 영상 데이타 표현 방법으로서 연구되어온 쿼드트리를 이용하여 한글 문서 영상을 저장하고, 쿼드트리로부터 문자들을 추출하였다.

각 문자는 문서 영상 분할 과정을 거쳐서 얻어진 윈도우에 의하여 추출하였으며, 이 때 심형 잡상도 자동으로 제거할 수 있었다. 분할은 3단계로 이루어지는데, 제 1단계에서는 인접한 검은 블럭들을 찾아서 명명(labeling)하고, 제 2단계에서는 제 1단계에서의 동치 쌍(equivalence pair)들의 목록을 가지고 동치 부류(equivalence class)를 형성한다. 이 때, 각 절편(segment)들은 절편의 위치 정보를 표현하는 윈도우를 생성한다. 제 3단계에서는 동치 부류들의 분류 번호를 갱신(update)함으로써 한 동치 부류 내의 동치들이 같은 분류 번호를 갖도록 한다. 이렇게 분할이 이루어지고 나면 절편들의 윈도우로부터 문자 추출을 위한 윈도우를 생성하고, 행과 열에 대하여 정렬(sort)된 절편들에 대하여 이 윈도우를 적용한다. 그 결과, 각 음절이 문맥상의 흐름 순서와 같이 아래 쪽의 문자보다는 위쪽의 문자가, 우측의 문자보다는 좌측의 문자가 먼저 추출되었다.

쿼드트리로 영상을 저장하므로 영상의 기억 장소가 현저히 감소하였으며, 이 쿼드트리로부터 직접 문자를 추출할 수 있었다. 본 연구에서는 동일한 크기의 문자만을 갖는 문서 영상을 대상으로 윈도우에 의한 추출이 이루어졌는데, 앞으로 문자 크기가 다른 것이 혼합되었다든지 그림이 혼합된 영상에서도 문자를 추출할 수 있도록 윈도우의 동적인 적용이 요구된다.

참고 문헌

[1] 이 인동, 권 오석, 김 태균(1991), "문서 인식을 위한 전처리 기술의 소개", 한국정보과학회 정보과학회지 제 9권 제 1호, pp. 14-21.

[2] H. Samet(1981), "Connected component labeling using quadtrees", *J. Assoc. Comput. Mach.* 28, pp. 487-501.

[3] H. Samet(1984), "The quadtree and related hierarchical data structure", *ACM Comput. Surveys* 16, pp. 187-260.

[4] E. Kawaguchi, T. Endo and J. Matsunaga(1983), "Depth-first picture expression viewed from digital picture processing", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. PAMI-5, No. 4, pp. 373-383.

[5] 장 선영(1990), "잡음 성분을 포함한 한글 문자 인식", 이화여자대학교 대학원 석사학위논문.

[6] 백 은경, 김 소연, 조 동섭(1991), "Quadtree 기법을 사용한 한글 폰트의 효율적인 압축에 관한 연구", 1991년도 하계 전자계산기연구회 학술연구발표회논문집.

[7] 남궁 재찬, 류 황빈, 남궁 윤(1988), "한국어 문서로부터 문자 분리 및 도형 추출에 관한 연구", 대한전자공학회 논문지, 제 25권 9호, pp. 1091-1100.