신문의 구조 분석을 통한 문자 영상의 추출

• 김 형 훈, 김 진 형

한국과학기술원, 전산학과

NEWSPAPER LAYOUT ANALYSIS and CHARACTER EXTRACTION

Hyunghoon Kim and Jin H Kim

KAIST, Department of Computer Science

요 약

이 논문은 스캐너로부터 얻어진 이진의 신문 영상을 분석하여 문자 단위의 영상을 추출하는 신문 인식시스템의 전처리 단계에 대한 연구이다 처리 시간을 줄이고 계속되는 작업을 쉽게 해주기 위해서 먼저 신문 영상을 블럭 리스트 표현으로 변환한다. 이 결과로부터 각 영역의 통계적 특성을 이용하여 신문의 구성 요소인 문자 영역, 그림 영역, 직선 영역 등으로 분류한다. 이를 영역의 상호 관계를 나타내는 그래프로 표현한다 그리고 신문의 구조를 반영한 정보 흐름에 대한 규칙을 이용하여 이 그래프로부터 한 기사가 어떻게 연결 되었나를 알아낸다. 이러한 일련의 작업을 거친후 문자 단위의 영상을 추출한다

I. 서 론

오늘날 대부분의 정보는 종이에 인쇄되어 전달되거나 저장된다 예를 들면, 사무실의 서류, 잡지, 교과서 등에 대한 중요한 매체로써 종이를 사용한다 전자 수단에의한 정보의 전달이 보다 실용적이고 값싸게 되자 사람들은 종이에 이미 기록된 정보를 이를 다루는 컴퓨터에 적절한 형태로 변환할 필요가 생겼다 사무 자동화에 있어서도 이러한 요구는 더욱 절실하게 요구되었다 사람이 키보드를 사용한 정보의 입력은 많은 비용이 들어간다. 광문자 인식기(OCR) 또한 미리 정해진 크기의 문자만을 인식할 수 있다 그러나 대부분의 문서는 문자와 그림을 혼용 하고있다

문서 자동 입력 시스템은 이러한 문자와 그림이 혼용된 문서를 자동적으로 컴퓨터에 입력하는 시스템이다 이 시스템은 전처리 단계와 문자 영상 인식 단계로 구분할 수 있다 전처리 단계에서는 문서 영상으로부터 문자 영상을 추출한다. 문자 영상 인식 단계는 추출된 문자 영상을 인식하여 문자 영상에 대한 코드를 얻는다 전처리 단계는 크게 영역 분할및 명시, 정보 흐름 결정, 문자 영상 추출의 3단계로 이루어진다. 문서는 크게 문자 영역과 문자가아닌 영역으로 나눌수 있으나 문서의 종류에 따라서 더욱더 세분할 필요가 있다 예를 들면, 신문의 경우에는 수직선과 수평선이 중요한 의미를 가지고 있기때문에 이들을 인식 하여야 한다 이와 같이 문서는 각각의 특성을 갖는 몇개의 영역으로 구성 된다 영역 분할및 명시단계에서는 이러한 영역을 블럭 형태로써 각 영역을 찾아낸다

이제 문서 영상에서 찾아낸 각 영역을 정보가 흐르는 순서대로 배열 해야한다 즉 수평으로 쓰여진 문장이 단순히 위에서 부터 아래로 배열된 문서의 경우에 정보가 흐르는 순서는 왓쪽에 있는 문장에서 아랫쪽에 있는 문장으로 배열만 하면 된다 그러나 신문이나 대부분의 보고서는 좀더 복잡한 구조를 가지고 있다. 이러한 문서에서 구해진 문장을 단순하게 왓쪽에서 아랫쪽으로 배열 한다면 이는 이 문서에 대한 옳바른 정보 흐름이 되지 못한다 따라서 복잡한 구조를 갖는 문서로 부터 옳바른 정보 흐름에 따른 문자 영상을 추출 하기 위해서는 그 문서의 특성에맞는 각 영역의 배열 원칙을 이용해야 한다 신문의 경우에는 규칙선이라는 영역이 정보의 흐름에 대한 안내자 역할을 한다. 한 문서의 정보 흐름을 결정하기 위해서는 문서에 배열된 영역들간의 위치 관계와 정보 흐름에 대한 안내자 역할을 하는 영역 그리고 기본적인 정보 흐름에 대한 경험적인 지식을 이용한다 기본적인 정보 흐름에 대한 경험적인 지식이란 각 문장이 수평으로 쓰여진 문서의 경우에는 그 문서의 왓쪽에 있는 문장에서 아랫쪽으로 정보의 흐름이 있다는 것과 각 문장이 수직으로 쓰여진 문서의 경우에는 그 문서의 오른쪽에서 왓쪽으로 정보의 흐름이 있다는

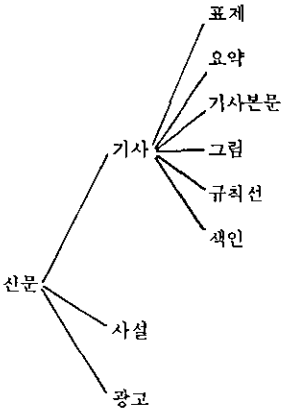
것을 말한다

이와 같은 정보 흐름이 정보 흐름 결성 단계에서 구해지고 그 흐름에 따라서 문자 영역으로부터 문자 영상을 추출한다 문자 영상 추출 단계는 문자의 크기 즉, 각 문자의 높이 또는 나비에 대한 정보를 주지 않아도 된다 문장 영상으로부터 문자 영상의 추출에 필요한 정보를 구한다 즉 문장의 높이로 추출할 문자 영상의 크기를 예측할 수 있다

이러한 기본적인 생각하에서 본 논문은 신문에 대한 전처리 단계를 구현 하였다. 논문의 구성은 다음과 같다 II장에서는 이논문이 다룰 신문의 여러 특성에 대해서 알아 보았다 III장에서는 영역 분할에대한 알고리즘을 기술하였고 IV장에서는 신문에서의 정보 흐름에 대해서 그리고 V장에서는 문자 영상의 추출에 대해서 다루었다

II. 신문의 구조분석

일반적으로 신문은 (그림 1)과 같은 몇개유형의 영역을 갖는다 각 영역은 블럭으로 나타내며 서로 겹치지 않는다



(그림 1) 일반적인 신문의 영역 유형

신문의 각 면은 잠정적으로 14개의 수평 영역으로 나누어진다 각 기사는 보통 여러개의 수평 영역에 존재한다 이때 각 수평 영역에 있는 조각을 기사본문조각이라 하자 각 영역의 통계적 특성은 다음과 같다 여기서 크기에 대한 단위는 스캐너에서 얻은 이진 영상의 화소를 기본 단위로 했다

(1) 표제 (a) 요약영역과 기사본문영역에 사용된 문자보다 더 큰 문자를 사용한다 (b) 밑바탕에 간단한 무늬를 갖는 경우도 있다 (c) 밑바탕에 무늬를 갖는 표제영역의 높이와 나비의 비가 3보다 크다

(2) 요약 : (a) 기사본문에 사용된 문자와 같은 크기의 문자를 사용한다 (b) 수직으로 쓰여진 경우에는 더 큰 높이를 가지며 수평으로 쓰여진 경우에는 더 큰 나비를 갖는다

(3) 기사본문 . (a) 사용된 문자의 높이와 나비는 각각 23화소의 크기를 갖는다 (b) 기사본문조각의 높이는 약 180화소의 크기를 갖는다

(4) 그림 : (a) 그림영역의 높이 또는 나비는 100화소 이상의 크기를 갖는다. (b) 그림영역의 높이와 나비의 비가 3보다 작다

(5) 규칙선 (a) 수평성분과 수직성분 두 종류가 있다 (b) 규칙선의 높이와 나비의 비가 보통20보다 크다

각 기사는 한개 이상의 표제영역, 요약영역, 여러개의 기사본문조각, 그리고 그림영역과같은 여러 유형의 영역이 여러 수평 영역에 존재한다 그러므로 신문으로부터 한 기사를 추출하기 위해서는 단지 그 기사에만 속하는 모든 영역을 찾아야 한다. 이를 위해서 신문의 구성 원리를 알아보자 이는 신문의 기사에 대해서만 적용된다

(1) 표제영역은 기사본문영역 시작부의 오른쪽 또는 윗쪽에 위치한다

(2) 요약영역은 기사본문영역과 표제영역 사이에 위치한다

(3) 기사본문 조각은 각 수평 영역 내에서는 오른쪽에서 왼쪽 그리고 윗쪽 수평 영역에서 아랫쪽 수평 영역으로 배열된다

(4) 기사본문 조각은 표제영역, 수직 규칙선, 그림을 가로질러서 배열되지 않는다.

III. 영역 명시및 분장 분할

3.1 이진 영상에대한 블럭 리스트 표현

처리 시간을 줄이고 계속되는 작업을 쉽게 하기위해서 이진영상을 연결된 화소의 블럭으로 표현한다. 문서 영상의 왼쪽 윗모서리를 원점(x = 0, y = 0)으로 하고 다음처리를 한다. 각 연결된 화소에 대해서 이를 포함하는 최소한의 블럭을 얻는다. 이 블럭은 그 블럭의 왼쪽 윗쪽 모서리(x_{min} y_{min})와 오른쪽 아랫쪽 모서리(x_{max} y_{max}) 를 나타내는 블럭(x_{min} y_{min} x_{max} y_{max})로 표현한다. 그 문서 영상에 존재하는 모든 연결된 화소에 대해서 이와 같은 블럭 표현을 얻는다. 만약 블럭(x_{min} y_{min} x_{max} y_{max})와 블럭 (x'_{min} y'_{min} x'_{max} y'_{max})이 조건 ($x_{min} \leq x'_{max}$, $x_{max} \geq x'_{min}$, $y_{min} \leq y'_{max}$, $y_{max} \geq y'_{min}$)을 만족하면 이 두 블럭은 블럭 (min{x_{min}, x'_{min}} min{y_{min}, y'_{min}} max{x_{max}, x'_{max}} max{y_{max}, y'_{max}})로 병합된다. 각 블럭은 문자영상, 규칙선, 그림등의 전체 또는 부분을 나타낸다.

3.2 영역 명시및 문장 분할

신문의 영역은 3.1절에서 분석한 유형의 영역을 갖는다. 이 절에서는 신문 이진 영상의 블럭 표현으로부터 이러한 영역을 찾는다. 3.1절에서 분석한 각 영역의 통계적 특성을 주로 이용 하였으며 알고리즘의 구현에서는 이러한 영역을 수직 성분과 수평 성분에 의해서 더욱 세분하여 각 영역을 찾는다. 신문의 구성 영역중에는 수직 규칙선과 수평 규칙선이 있는데 이 규칙선에 의해서 신문은 여러 영역으로 나누어진다. 이러한 정보를 이용하여 이 알고리즘의 속도를 높일 수 있으며 또한 보다 정확하게 영역을 분할하게 된다. 수직 규칙선은 신문의 구성영역중 수평 성분을 갖는 영역들의 존재 구역을 대충 결정해주고 수평 규칙선은 수직 성분을 갖는 영역들의 존재 구역을 결정해준다. 수직 성분을 갖는 구성 영역은 수직 표제영역, 밑바탕을 갖는 수직 표제영역, 기사 본문 조각, 수직 규칙선, 수직 색인이며 수평 성분을 갖는 구성 영역은 수평 표제영역, 밑바탕을 갖는 수평 표제 영역, 수평 규칙선, 수평 색인이다. 그림 영역은 아무 성분도 갖지 않는다. 각 영역을 명시

하는 순서는 수평 규칙선 영역, 수직 규칙선 영역, 기사 본문 영역, 밑바닥을 갖는 수직 표제 영역, 밑바탕을 갖는 수평 표제 영역, 그림, 수평 표제 영역, 수직 표제 영역, 그리고 수직 색인, 수평 색인 순으로한다. 이 알고리즘에 대한 입력은 이진영상에대한 블럭표현으로 한다. 각 영역 R_i유형을 명시하는 알고리즘을 간략하게 기술하면 아래와같다.

(a) 영역 R_i유형일 가능성이 있는 블럭을 모은다.
(b) 수직 규칙선 또는 수평 규칙선이 명시 되었다면 이들에 의해서 모아진 블럭을 여러 영역으로 나눈다.
(c) 나누어진 각 영역의 블럭에 대해서 블럭 간의 간격을 조사하여 영역R_i유형의의 한 영역에 포함될 가능성이 있으면 한 블럭으로 병합한다.
(d) (c)에 의해서 구성된 영역이 영역R_i유형의 특성을 만족하는지 조사한다.
(e) 만약 문자 영역이면 수평 투영 또는 수직 투영에 의해서 각 문장을 분할한다.

IV. 기사영역의 추출

4.1 신문의 구조분석과 그래프 표현

신문 영상으로부터 기사를 추출하기위해서는 신문에 존재하는 각 영역의 배열 형태를 알아야한다 배열 형태는 신문에 나열된 각 영역 간의 위치 관계에 의해서 알 수 있고 위치 관계란 각 영역의 상하 좌우에 근접성을 만족 한면서 존재하는 영역에 의해서 표현된다 그래프의 vertex는 각 영역을 나타내며 edge는 상하 좌우에 근접성을 갖는 영역에 대한 정보를 갖는다 영역 R_i와 R_j가 상하 위치 관계로 근접성을 가질 때 이를 수직 근접성이라 하며 좌우 위치 관계로 근접성을 가질때는 수평 근접성을 가진다고 한다 이들의 정확한 정의는 생략한다.

4.2 정보 흐름 결정

3.2절의 영역 분할및 문장 분할 단계에 의해서 신문 영상에 있는 모든 구성 영역을 찾았다 하지만 찾아낸 이러한 구성 영역을 어떤 순으로 배열 해야 되는지 알지 못한다 더구나 신문의 한 면에는 여러 기사가 존재하므로 각 기사별로 문자 영상을 추출 해야한다 이제 정보 흐름에 대한 규칙을 4.1절의 구조 분석에 의해서 얻어낸 그래프에 적용 하여 각 기사를 추출한다 각 기사는 한개 이상의 표제 영역을 갖고 있으며 표제 영역의 왼쪽에 기사의 시작 기사 본문 조각이 있다는 사실을 이용한다. 따라서 먼저 표제 영역을 찾고 이 표제 영역의 왼쪽에서 기사 본문 조각의 시작부를 찾는다 각 기사 본문 조각은 정보 흐름에의해서 찾는다 정보 흐름 결정에 의한 기사 추출 알고리즘은 다음과 같이 기술된다

(a) 표제영역을 찾는다
(b) 만약 표제영역이 없으면 이 작업을 마친다
(c) 만약 표제영역이 있으면 표제영역을 문자영상 추출단계로 보낸다
(d) 표제영역을 이용하여 기사본문의 시작부를 찾는다 이를 cur-text라 하자
(e) cur-text를 문자영상 추출단계로 보낸다
(f) cur-text를 입력으로 하는 다음 기사 본문 조각 결정 루틴 (next-text를 찾는다)
(g) 만약 next-text를 발견 할 수 없으면 현재 기사의 종료를 알리고 (a)로 간다
(h) 만약 next-text를 발견 하였으면 cur-text의 값을 next-text로 치환하고 (e)로 간다

cur-text로 부터 정보흐름 규칙을 이용하여 next-text를 결정하는 다음 기사 본문 조각 결정 알고리즘은 다음과 같다

(a) cur-text의 왼쪽에 근접한 영역을 조사한다
(b) 만약 기사본문 조각이 있으면 next-text로 이 기사본문 조각을 출력한다.
(c) 만약 그렇지않으면 cur-text의 아랫쪽에 있는 영역을 조사한다
(d) 만약 수평 규칙선이 있으면 수평 규칙선의 아랫쪽에 있는 영역중에서 제일 오른쪽에 있는 영역을 조사한다.
(e) 제일 오른쪽에 있는 영역이 기사본문이 아닌 다른 영역이면 현재 기사의 종료를 알린다.
(f) 만약 기사본문 조각이면 cur-text와 이 기사본문 조각의 정보흐름 상태를 결정한다.
(g) 연속되는 정보흐름 상태를 갖는다면 next-text로서 출력한다.
(h) 불연속 상태를 나타내면 현재 기사의 종료를 알린다.

cur-text와 next-text의 정보흐름 상태의 결정은 다음과 같다 cur-text의 제일 왼쪽 문장의 아랫쪽에 문자가 몇개 비어있고 next-text의 제일 오른쪽 문장의 윗쪽에 문자가 몇개 비어 있거나 cur-text의 제일 왼쪽 문장의 아랫쪽에 문자가 가득 채워져 있고 next-text의 제일 오른쪽 문장의 윗쪽에 문자가 가득 채워져 있을때는 cur-text와 next-text는 연결성을 갖는다.

V. 문자 단위 영상의 추출

이 단계는 문장으로 부터 문자 단위 영상을 추출한다 윗단계의 처리에 의해서 구별된 수직 문장과 수평 문장으로부터 문자 영상을 추출한다 기사본문의 문장에서는 일정한 크기의 문자를 사용하기때문에 간단하게 문자 영상을 추출 할 수 있다 그러나 표제영역의 문자는 일정한 크기의 문자를 사용 하지 않기 때문에 문제가 된다 그래서 본 논문에서는 각 문장의 높이에 대한 정보를 이용하여 문자 영상을 추출한다 먼저 문장 영상에 대한 투영을 계산한다 만약 수평 문장이면 수직 방향으로 투영 하고 수직 문장이면 수평 방향으로 투영한다 이 투영에된 결과에서 어떤 임계치보다 큰 투영 결과를 갖는 영역의 시작 장소와 끝나는 장소의 리스트를 구한다 이 리스트를 투영 리스트라 하자 어떤 임계치보다 큰 투영 결과를 갖는 부분에 문자 영상의 전체 또는 문자 영상의 부분이 존재할것이다. 투영 리스트로 부터 어떤 임계치보다 큰 부분의 길이를 구할 수 있으므로 이 값과 문장의 높이를 비교하여 문자 영상을 추출할 수 있다 문장의 높이로 그 문장에 포함된 문자 영상의 크기를 예측할 수 있다

VI. 결론

신문 자동 입력 시스템의 전처리 단계에 대해서 논 하였다 신문 영상의 블럭 표현을 사용함으로써 화소 단위 처리에서 연결된 화소 단위로 처리하였다 신문 영상의 영역중에서 그림 영역과 같은 경우에는 그림 전체가 한개의 연결된 화소

가 되므로 문자 단위 영상 추출이 필요 없다 따라서 처리 속도를 높일 수 있다. 영역 분할및 문장 분할에서는 각 영역의 높이와 나비에 대한 통계적 특성을 이용 하였으며 수직 규칙선과 수평 규칙선을 제일 먼저 찾아서 다른 영역을 찾을때 이를 이용하였다 이 들을 이용함으로써 처리 속도를 높이고 보다 정확한 분할을 한다. 기사의 추출은 문서에서의 각 영역의 상호 관계, 정보 흐름에 대한 규칙 그리고 정보 흐름을 제어하는 영역들을 이용하였다. 문자 영상의 추출은 문장의 높이로 문자 영상의 크기를 예측하여 문자 영상을 추출하였다 본 논문에서는 스캐너의 받아 들일수 있는 문서의 크기가 A4 용지보다 작은 것이어야 하고 또 사용하는 컴퓨터의 기억장치의 한계 때문에 신문 한 면의 A4 용지 크기만을 처리한다 현재 구현된 시스템에서는 신문 기사에 대해서는 양호한 결과를 얻는다. 즉, 각 기사를 추출하고 그 기사에 대한 문자 영상을 추출한다 그러나 밑바탕에 무늬를 갖는 표제 영역에 존재하는 문자 영상의 추출 문제가 해결되지 않았다 또한 사실, 광고 영역의 처리 문제가 남아 있다 이 시스템은 LISP 언어로 구현 하였다.

참 고 문 헌

[1] J. Toyoda, Y. Noguchi, Y. Nishimura, "Study of Extracting Japanese Newspaper Article," IEEE, 1982

[2] Shoji Ito, Shinji Sakatani, "Field Segmentation and Classification in Document Image," IEEE, 1982

[3] K Y Wong, R. G. Casey, "Document Analysis System," IBM J Res Develop , Vol 26, No 6, November 1982

[4] Kosaku Inagaki, Toshikazu Kato, Tadashi Hiroshima, Toshiyuki Sakai, "MACSYM A Hierarchical Parallel Image Processing System for Event-Driven Pattern Understanding of Documents," Pattern Recognition, Vol 17, No 1, pp. 85 - 108, 1984