論文 97-22-4-12

패턴매칭에 의한 이진 한글문서의 유·무손실 압축에 관한 연구

正會員 김 영 태*, 고 형 화*

The study on lossy and lossless compression of binary Hangul textual images by pattern matching

Young Tae Kim*, Hyung Hwa Ko * *Regular Members*

※본 논문은 통상산업부 연구비지원에 의한 결과임

요 약

패턴매칭에 의한 문서의 압축은 추출한 패턴간의 상관성을 개발하여 부호화하는 방식이다. 패턴매칭에 의한 한글 문서를 압축할 경우에 조합형인 한글의 특성상 불규칙한 자소의 접촉으로 인하여 패턴간의 상관성이 떨어진다. 따라서 본 논문에서는 자소단위의 매칭을 유도하여 패턴간의 상관성을 효율적으로 이용하기 위하여 자소분리를 시도하였다. 자소분리는 추출된 패턴의 모음 유무를 판별하여 모음과 연결된 자음의 접촉점을 찾아 분리하는 방법을 사용하였다. 제안한 알고리즘을 기존의 방법과 비교했을 때 압축율이 유손실 모드인 경우에는 기존의 PMS[5]보다 3.4%- 9.1%, 무손실인 경우에는 차세대 이진 영상 압축알고리즘으로 표준화 위원회에 상정된 SPM [7]의 무손실 모드 보다 1.3%-3.0% 향상되었다.

ABSTRACT

The textual image compression by pattern matching is a coding scheme that exploits the correlations between patterns. When we compress the Hangul (Korean character) text by pattern matching, the correlations between patterns may decrease due to random contacts between phonemes. Therefore in this paper we separate connected phonemes to exploit effectively the correlation between patterns by inducing the match. In the process of separation, we decide whether the patterns have vowel component or not, and then vowels connected with consonant are

*광운대학교 전자통신공학과
論文番號:97075-0224
接受日字:1997年 2月 24日

separated. When we compare the proposed algorithm with the existing algorithm, the compression ratio is increased by 1.3%-3.0% than PMS[5] in lossy mode, by 3.4%-9.1% in lossless mode than that of SPM[7] which is submitted to standard committee for second generation binary compression algorithm

Ⅰ. 서 론

영상정보의 방대함으로 인한 압축의 필요성과 통신을 위한 표준화의 필요성이 제기되어 압축알고리즘의 국제적인 표준화가 구축되었다. 이러한 표준화와 통신환경의 계속적인 변화는 다른 응용분야가 창출되도록 자극하였고 창출된 새로운 응용분야는 또 다른 표준화를 요구하게 되었다. 영상압축 분야에서도 새로운 응용분야에 대한 수요를 충족시키기 위한 차세대 표준압축 알고리즘이 상정되기에 이르렀다. 국제표준화기구(ISO/IEC JTC1 SC29)에서는 구체적으로 정지영상을 대상으로 이진영상의 유·무손실 부호화, 다계조 영상의 무손실 압축, 다중 성분 압축 그리고 5차원영상 압축등 4개항을 상정하여 표준안 개발작업을 진행중에 있다. 한편 영상압축의 다양한 응용분야가 각기 독립된 시스템을 요구한다면 이는 경제적으로 비효율적이다. 따라서 응용영역이 확대되지만 이를 하나의 시스템에서 처리할 수 있는 통합된 알고리즘에 대한 필요성도 제기되어 새로운 프로젝트로 JPEG-2000을 상정하게 되었다[1].

이진(bi-level)영상 압축분야를 살펴보면 그 동안 구축된 표준안들은 모두 무손실(lossless)압축이다. 응용에 따라 화소단위의 정보손실(lossy)을 허용하면 무손실기법이 제공하는 압축율의 한계를 뛰어넘을 것이라는 취지에서 유손실 모드의 도입이 제기된 것으로 사료된다. 기존의 표준화된 이진(bi-level)영상 압축알고리즘은 MR, MMR, JBIG등이 있다. 유손실은 주로 텍스트 문서를 대상으로 패턴매칭(pattern matching) 기법을 사용하여 연구되었다. 패턴매칭에 의한 문서의 압축은 Ascher[4]등에 의해 1974년 처음으로 제안되었고 지금까지 꾸준히 연구되고 있다[5-7]. 특히 Howard에 의해 제안된 논문은[7] 산술부호화와 패턴매칭을 근간으로 하고 있으며 국제 표준화 위원회(ISO/IEC JTC1 SC29/WG1)에서 추진하고 있는 차세대 이진영상의 유·무손실 부호화의 표준안(JBIG-2)으로 상정되었다. 이러한 기존의 연구되어진 패턴매칭기법에 의한 문서의 압축은 비교적 우수한 성능을 보이지만 한글문서에 그대로 적용하면 영문자와 달리 조합형인 한글은 그 특성상 자소단위의 접촉이 다양하게 발생하므로 추출된 패턴간의 상관성이 떨어져 패턴매칭의 효과가 떨어진다. 따라서 본 논문에서는 전처리 단계로서 추출된 패턴의 자소의 접촉점을 찾아 분리해내는 방법을 사용하여 자소단위의 매칭이 이루어지도록 유도함으로 패턴간의 상관성이 효율적으로 개발되도록 시도하였다. 자소분리는 패턴인식분야에서 자소단위의 인식을 위해 연구의 필요성이 제기되어 연구되어진 바 있다[8, 9]. 본 논문에서 사용한 자소분리 과정은 자소가 접촉될 가능성이 있는 패턴을 대상으로 먼저 모음성분의 존재 여부를 판별한 후 모음축과 연결된 가지점(branch node)을 찾아 이 점이 모음 자체의 것인지 아니면 자음과의 불규칙한 접촉으로 발생된 것인지를 판별하게 된다. 후자인 경우에 분리위치를 찾아 자소가 분리되어 추출되도록 한다. 본 논문의 구성은 다음과 같다. 2장에서는 기존의 연구되어 온 알고리즘을 살펴보고 3장에서는 패턴매칭에 의한 한글문서 부호화의 성능개선을 위하여 제안된 자소분리과정을 면밀히 살펴보고 4장에서는 모의실험에 대한 결과와 고찰을 다룬다. 끝으로 5장에서 결론을 이끌어 내었다.

Ⅱ. 이진영상 압축알고리즘

이진영상 압축알고리즘은 크게 무손실(lossless)과 유손실(lossy)알고리즘으로 나뉜다. 최근에는 유·무손실 모드를 모두 제공하는 알고리즘이 발표되었다[6, 7].

1. 무손실 압축(lossless compression)

무손실 압축으로는 G3, G4 팩시밀리에 사용된 표준 알고리즘을 들 수 있다[2]. 이보다 진보된 형태로 확률적인 모델에 근거해서 비트열을 최적으로 부호화하는 산술부호화(arithmetic coding)를 들 수 있다. 이는 송·수신단에서 동시에 이용할 수 있는 이전의

부호화된 화소값을 콘텍스트(context)로 하여 현재 부호화 할 화소의 확률값을 산출하여 부호화한다. 산술부호화는 JBIG(Joint Bi-level Image Experts Group) 표준알고리즘의 근간이 되는 것으로 10개의 화소로 템플레이트(template)를 구성하여 콘텍스트로 사용한다.

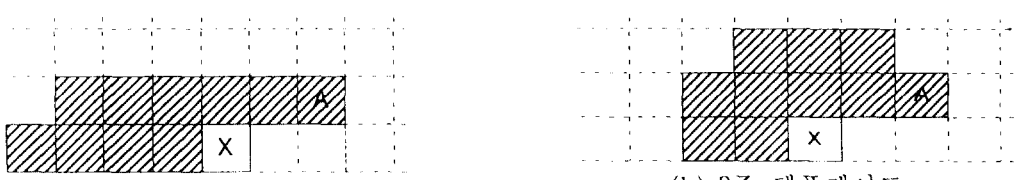
2. 유손실 압축(lossy compression)

유손실 기법은 텍스트 문서를 대상으로 패턴매칭(patten matching)을 근간으로하는 알고리즘이 그동안 연구되었다[4-7]. 패턴(pattern)이란 연속되어진 흑화소의 집합으로 정의 할 수 있다. 패턴매칭(pattern matching)은 패턴에 관한 라이브러리(library)를 구성하여 반복적으로 나오는 패턴의 비트맵(bit map)을 일일이 부호화하지 않고 매칭패턴에 관한 라이브러리 인덱스를 부호화하는 방식이다. 부호화할 패턴과 라이브러리에 있는 패턴의 중심점을 잡아 일치되도록 겹쳐놓고 배타적 논리합(exclusive -OR)을 계산하여 에러 비트맵(bit map)을 만든다. 에러 비트맵에서 "1"의 비율이 미리 트래이닝을 통해 결정된 임계치보다 크면 상이한 패턴으로 그렇치 않으면 동일패턴으로 판정하여 처리한다. 텍스트 문서 압축을 위한 기본적인 착상은 Ascher[4]등에 의해 1974년에 처음으로 제안되었다. Johnsen[9]등은 PMS(Pattern Matching and Substitution)라 불리우는 알고리즘을 제안하였다. 인덱스 정보를 효율적으로 부호화하기 위해 라이브러리 안에 놓여진 패턴의 순서가 최근에 사용되어진 빈도수에 근거하여 재배치되도록 구성하였다. 또한 잘못 매칭되어진 것을 줄이기 위해 에러비트맵에서 가중치를 두어 에러율을 계산하는 방법도 제안하였다.

3. 유·무손실 압축

최근에는 유손실과 무손실을 응용에 따라 선택적으로 사용할수있는 알고리즘이 제안되었다. Witten[6] 등은 1단계로 유손실을 2단계로 무손실을 제공하는 알고리즘을 제안하였다. 1단계로 패턴을 추출하여 부호화하고 2단계로 잔여(residue)영상을 부호화한다. 잔여영상은 라이브러리를 구성할 때 제외된 패턴과 완전한 매칭이 되지 않아 주로 에지 부근에 발생하는 에러등에 관한 정보가 된다. 잔여영상에 해당하는 정보의 부호화는 원영상과 수신단에서 이미 유손실 모드로 복원된 영상이 있으므로 이 복원영상을 결합하여 템플레이트를 구성하여 산술부호화한다. 이때 복원된 영상에 관한 템플레이트는 그림2(a)와 같이 부호화할 화소와 인접한 화소를 모두 이용하여 구성(clairvoyant template)할 수 있다. 이것이 그림1과의 차이점이다. 그림2(a)와 같은 템플레이트를 구성하여 예측도를 높여 부호화 할 수 있다.

Howard[7]는 소프트 패턴매칭(SPM:Soft Pattern Matching) 알고리즘을 제안하였다. 패턴을 추출하여 라이브러리를 탐색하여 새로운 패턴으로 판정되면 이 패턴을 JBIG에서 사용한 템플레이트를 이용하여 마치 하나의 이미지처럼 부호화한다. 매칭된 패턴의 비트맵을 부호화하는 방법은 그림2(b)와 같이 라이브러리에 있는 매칭 패턴의 화소와 현 패턴의 이미 부호화된 화소값을 템플레이트로 구성하여 산술 부호화한다. 그림2(b)에서 현재패턴의 "X"로 표기된 화소와 매칭패턴의 "7"로 표기된 화소가 서로 일치되도록 두 패턴이 겹친다. 이 알고리즘은 차세대 이진영상 표준 압축 알고리즘(JBIG-2)으로 제안되었다. 매칭된 패턴의 비트맵도 부호화하므로 무손실 모드가 되어



(a)2줄 템플레이트 (b) 3줄 템플레이트

(X : 부호화할 화소, A: 적응적인 화소)

그림 1. 템플레이트 화소(template pixels)

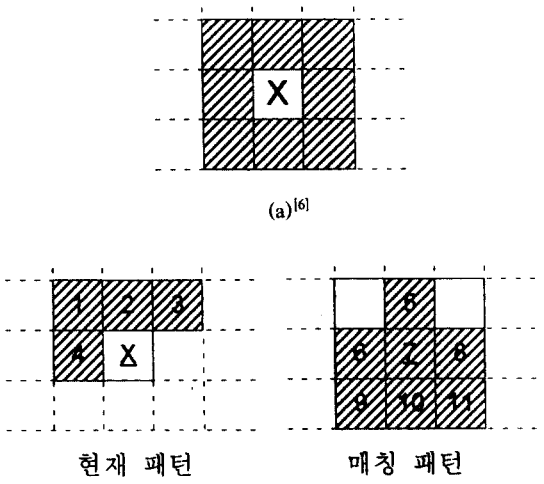


그림 2. 잔여영상 부호화를 위한 템플레이트 화소

기존의 패턴매칭 알고리즘에서 문제가 되었던 오매칭을 해결하게 되었다. 이 알고리즘은 유손실 모드로의 확장에 대해서도 언급하고 있다. 영상을 처리하기 전에 전처리로써 이미지 향상(enhancement) 기법을 사용하면 부호화 비트수도 감소되는 효과가 있다. 잡음제거라든가, 에지를 매끄럽게(smoothing)하는 기술이 그 예가 될 수 있다. 부호화 비트수를 줄이기 위해 특별히 고안된 알고리즘을 적용할 수도 있다. 매칭된 패턴의 비트맵을 부호화 할 때 라이브러리에 있는 패턴과 오매칭된 화소가 고립되어 있으면 이 화소값을 부호화하기 전에 바꿈(reverse)으로써 부호화 비트수를 줄일 수 있다. 또한 이 알고리즘은 점진적인 유,무손실모드도 제공한다. 이는 앞에서 언급한 유손실모드로 먼저 부호화하고, 유손실모드 부호화과정에서 적용된 알고리즘에서 바뀐화소값에 대한 정보를 다시 부호화하게 된다. 부호화해야 할 값이 패턴의 비트맵, 매칭여부등과 같이 이진(binary)인 경우에는 JBIG에 사용된 산술부호화(QM-coder[12])를 사용하여 부호화하였고 인덱스(index), 패턴의 위치, 크기정보와 같이 다차원인 경우에는 산술부호화를 다차원으로 확장한 방법(multisymbol QM-coder[7])을 사용하여 부호화하였다.

Ⅲ. 자소분리를 통한 한글문서의 효율적인 부호화

이진영상 압축알고리즘은 2장에서 살펴본 것과 같이 패턴매칭과 산술부호화를 근간으로 하는 알고리즘[5, 7]이 성능도 우수하고 유·무손실 모드를 제공하는 등 여러가지 장점을 갖고있다. 그러나 기존의 발표되어진 이러한 알고리즘을 한글에 그대로 적용할 경우 그 성능이 저하될 수 있다. 한글은 영문자와 달리 자음과 모음의 조합으로 문자가 구성되는 조합문자라는 특징을 갖는다. 따라서 자소간의 불규칙한 접촉으로 말미암아 자소의 수보다 훨신 많은 다양한 패턴들이 발생되어 추출된 패턴간의 상관도(correlation)가 떨어지게 된다. 본 논문에서는 자소의 불규칙한 접촉

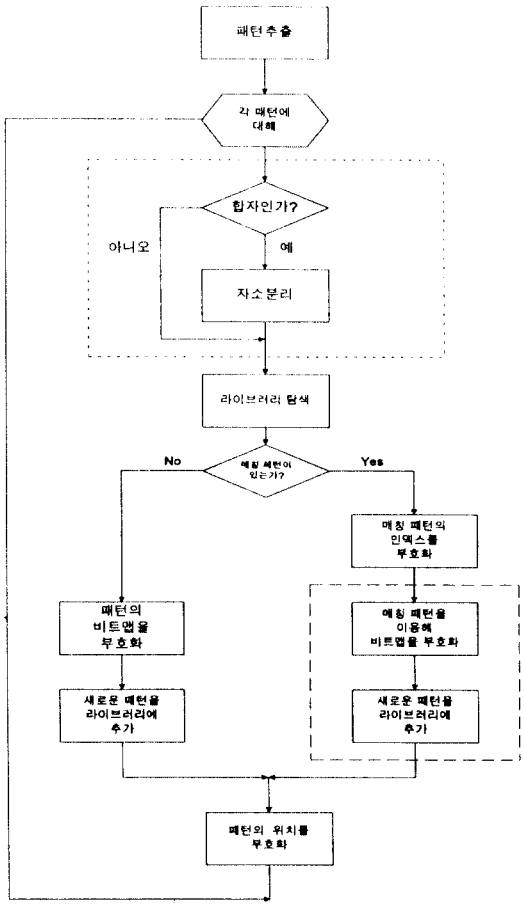


그림 3. 한글문서의 유·무손실 압축 알고리즘 흐름도

으로 인하여 생긴 합자(ligature)를 분리함으로 실질적으로 자소단위의 매칭을 유도하여 패턴간의 상관성이 효율적으로 개발되도록 시도하였다. 자소분리는 기존의 문자인식 영역에서도 자소단위로 인식하기 위하여 연구되어진 바 있다[8, 9]. 남궁재찬[8]은 인쇄체를 대상으로 종모음과 횡모음, 종성 유무를 판별하여 한글을 5형식으로 분류하였고 인덱스를 부여하여 창(window)의 크기를 결정하여 전처리로서 자소분리를 하였다. 최필웅[9]등은 필기체를 대상으로 가지점에서 국소 그래프패턴을 구성한 후 주가지와 부가지의 방향과 특징점 정보를 이용하여 자소접촉점을 찾아 분리하는 알고리즘을 제안하였다. 본 논문에서는 먼저 추출된 패턴이 합자일 가능성이 있는 후보를 선택한 후 모음의 유·무를 판정하여 모음에서의 가지점이 접촉으로 인하여 발생한 것인지 아니면 모음 자체의 성분인지를 판별하여 분리하는 방법을 사용하였다. 본 논문에서 제안한 자소분리 방법은 참고문헌 [13], [14]을 근간으로 하여 본 상황에 맞게 적응적으로 변형시켰다. 제안한 알고리즘의 전체 흐름도를 살펴보면 그림4와 같다. 기존의 알고리즘[12]에 짧은 점선으로 표시된 부분 즉 자소분리단을 전처리로써 추가하였다. 자소분리단 이외의 모듈은 기존의 방법을 유지하고 있다. 길은 점선으로 표기된 부분이 생략되면 유손실 모드이고 그렇지 않으면 무손실 모드가 된다.

1. 합자의 후보문자 산출

문서의 스캐닝과정에서의 왜곡 또는 폰트의 자체 모양에 기인하여 자소가 접촉된 패턴(합자)이 다수 나오게 된다. 모든 추출된 패턴을 대상으로 자소분리를 시도하지 않고 자소가 접촉될 가능성이 있는 것들을 후보로 판정하여 이것을 대상으로 한다. 합자는 그 크기가 일정정도 이상일 것으로 생각하여 합자의 크기를 추정하여 추정된 크기 범위에 있는 패턴말을 후보로 선정하였다. 합자의 크기추정은 히스토그램의 주기적인 특성을 이용하였다.

2. 자소분리

자소분리과정은 그림4과 같다. 먼저 세선화를 거친 후 특징점을 찾는다. 다음에 모음의 유무를 판별하고 모음에있는 가지점이 자소접촉점 인지를 결정한 후 분리위치를 지정하여 분리한다.

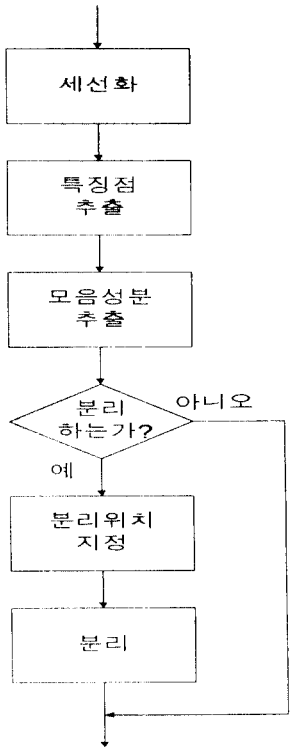


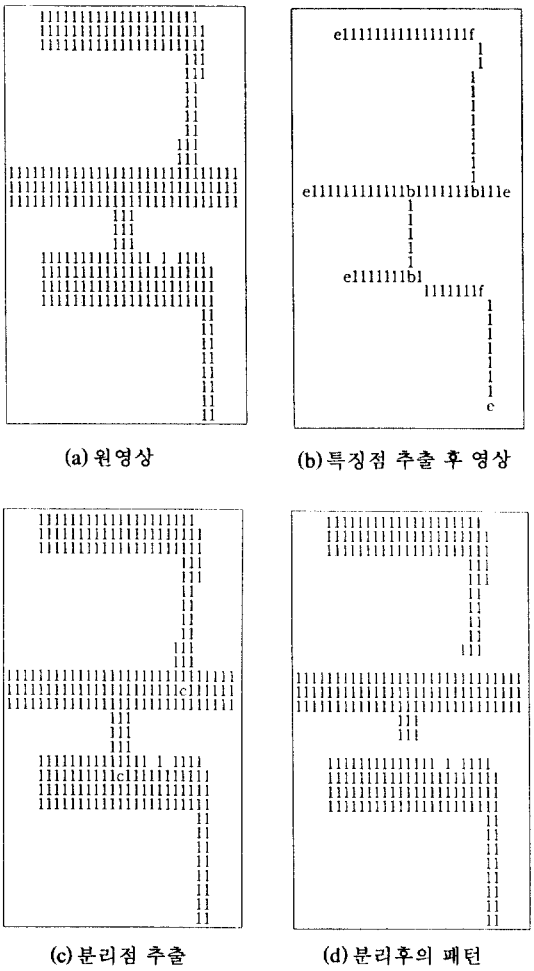
그림 4. 분리과정의 흐름도

① 세선화

세선화(thinning)는 입력패턴으로부터 골격선(skeleton)이라 불리는 형태로 변형시키는 것을 말한다. 본 논문에서는 chu[18]등에 의해 제안된 알고리즘을 사용하였다.

② 특징점 추출

특징점은 세선화된 패턴에서 일반적인 흑화소와 구별되는 특징을 갖고 있는 화소를 말한다. 여기에서 다루는 특징점은 끝점(end point), 굴곡점(fork point), 가지점(branch point)등이다. 끝점은 이웃하고 있는 화소가 한개인 경우이고 굴곡점은 2개이면서 서로의 방향이 수평하지 아니한 경우이고 가지점은 이웃하고 있는 화소가 3개인 경우이다. 그림5는 "국"이라는 패턴의 분리과정을 보여주고 있는데 (a)는 원영상이고 (b)는 세선화와 특징점 추출을 한후의 영상이다.



(a) 원영상 (b) 특징점 추출 후 영상

(c) 분리점 추출 (d) 분리후의 패턴

그림 5. 각 단계에서의 패턴의 형태

③ 모음성분 판별

종모음추출은 그림9(a)와 같이 추출된 문자의 우상단 1/4 영역을 탐색하여 끝점을 찾아 아래 방향으로 추적하여 그 길이가 추출된 패턴길이의 반이상이면 종모음성분이 존재하는 것으로 결정한다. 횡모음도 종모음과 유사한방법으로 유·무를 판별한다. 패턴의 왼쪽에 있는 끝점을 찾아서 오른쪽으로 추적하여 가지점을 만나고 가지점 이후의 길이가 패턴폭의 일정비율 이상일때 횡모음의 존재로 판정한다.

④ 가지점과 연결된 특징점에 의한 접촉유형의 사전 구성

횡모음과 종모음의 유무를 파악하여 다음과 같이 4가지로 나누어 각각의 경우에 독립적으로 접촉유형을 산출해 낸다.

1) 횡모음과 종모음의 접촉유형

종모음과 횡모음이 모두 존재하는 패턴으로 판정된 경우에는 종모음과 횡모음의 접촉점을 찾도록 한다. 분리된 패턴은 또 하나의 독립된 패턴으로 간주되어 다시 처리된다. 즉 분리된 패턴의 하나는 횡모음만 다른 하나는 종모음만 있는 패턴으로 처리될 것이다. 그림6과 같이 종모음의 끝점(E)에서 아래(down) 방향으로 추적하여 첫번째 가지점에서(B)에서 왼쪽길이를 측정하여 이것이 일정길이 이상이면 B점의 왼쪽부분을 분리해 낸다. 여기서 E는 끝점을 B는 가지점, F는 굴곡점을 X는 그 특징점이 무엇이든 고려하지 않는(don't care)점을 가르킨다. "과", "와" 등을 그 예로 들 수 있다.

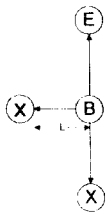


그림 6. 횡모음과 종모음의 접촉 유형

2) 횡모음과 연결된 자음의 접촉 유형

횡모음이 있을 경우에 횡모음과 초성, 횡모음과 종성이 접촉되는 경우가 있다. 그림7의(a), (b)는 횡모음과 초성이 접촉되는 경우이고 (c), (d)는 횡모음과 종성이 접촉되는 경우를 나타낸 것이다. 그림7(a)는 가지점 B에서 위(up)방향으로 추적하였을 때 굴곡점이 나오는 경우로 "고"자등을 그 예로 들 수 있다. 그림12(b)는 가지점B1에서 위(up)방향으로 추적하여 가지점 B2가 추출되면 가지점 B2와 B1-B2선분을 분리한다. "오", "노", "로" 등이 그 예가 된다. 그림7(c)과 같이 가지점B에서 아래(down)방향으로 추적하여 굴곡점이 추출되면 가지점의 아래 부분을 분리해 낸다. "는"등이 그 예가 된다. 그림12(d)와 같이 가지점B1

에서 아래(down)방향으로 추적하여 가지점B2가 추출되면 가지점 B2의 윗부분을 분리해 낸다. "국"에서 "ㅜ"와 종성 "ㄱ"의 접촉을 그 예로 들 수 있다.

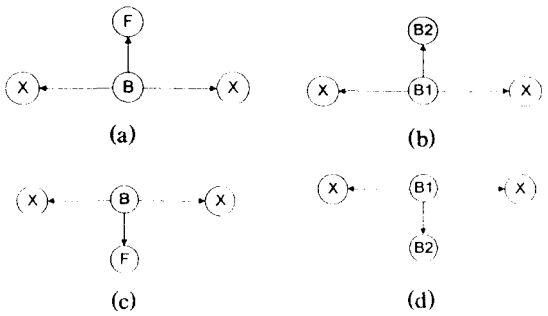


그림 7. 횡모음과 연결된 자음의 접촉 유형

3) 종모음과 연결된 자음의 접촉 유형

종모음과 초성또는 종성이 접촉되는 경우이다. 종모음이 이중모음인 경우에는 바깥쪽(오른쪽)에 있는 모음축의 가지점을 찾아 처리하게 된다. 따라서 가지점에서의 추적은 왼쪽으로만 시도한다. 가지점에서 아래로 추적하여 끝점이 나오면(그림8(a), (b), (c))이는 일단 종성이 없는 경우로 판정된다. 따라서 오른쪽으로 추적하여 만나는 특징점에 따라 초성과의 접촉여부가 결정된다. 그림8(a)처럼 굴곡점을 만나면 가지점의 오른쪽을 분리점으로 결정된다. 이같은 경우는 "너"를 예로 들 수 있다. 오른쪽으로 추적하여 그림8(b)처럼 가지점B2을 만나면 B2의 오른쪽을 분리한다. "어"자 등이 그 예가 된다. 오른쪽으로 추적하여 끝점이 나오면(그림8(c)) 끝점까지의 길이(L1)를 계산하여 이것이 패턴의 폭(Width)의 1/2 이하이면 분리점이 없는 경우이고, 이상이면 다시 L2의 길이를 계산한다. L2의 길이가 일정길이 이상이면 가지점의 위쪽을 분리한다. 그렇지 않으면 B의 왼쪽부분이 분리위치가 된다. 전자는 "넉" "덕"자등의 종모음과 종성의 접촉이 이 경우에 해당되며 후자는 "의"자가 여기에 해당된다. 가지점에서 아래로 추적하여 그림8(d)처럼 굴곡점이 추출되면 이때의 가지점은 종모음과 종성과의 접촉으로 인한 가지점으로 판정하여 가지점의 윗부분을 분리한다. "결"자에서 "ㅕ"와 "ㄹ"의 접촉이 그 예가 된다. 그림8(e), (f)에서 처럼 가지점의 아래 부분을 추적하여 다시 가지점(B2)이 나오면 가지점B1은 초성과의 접촉 여부를 가리게 된다. 즉 그림8(e)처럼 왼쪽의 특징점이 굴곡점이면 B1의 왼쪽부분을 분리위치로 그림8(f)처럼 다시 가지점이면 B3의 오른쪽을 분리위치로 결정한다. 끝으로 가지점에서 아래로 추적하여 특징점을 만나기 전에 좌우로 일정길이 이상 치우치면 추적을 멈춘다. 이는 "o"과 같은 순환 패턴을 만났음을 의미한다. "경"자 같은 경우이다. 이런 경우는 B의 윗부분이 분리위치가 된다. 그림8(g)가 여기에 해당된다. 특징점을 만나기 전에 추적이 끝나므로 흑화소를 의미하는 "1"로 표기했다.

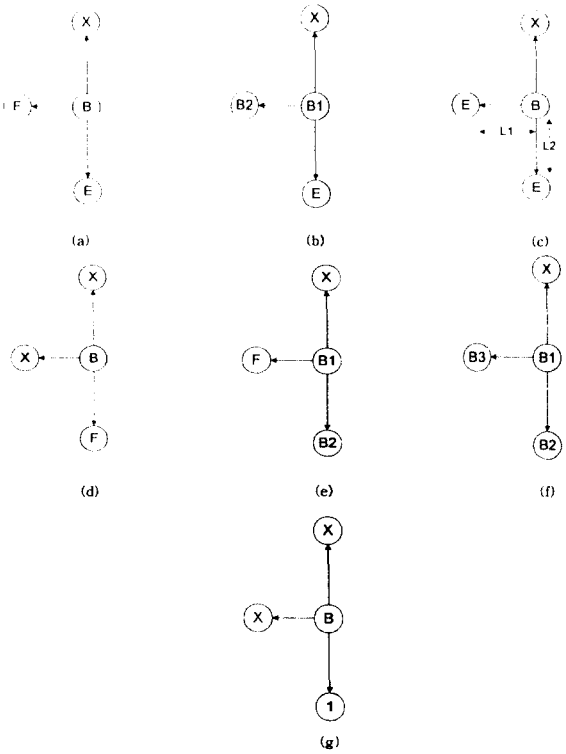


그림 8. 종모음과 연결된 자음의 접촉 유형

4) 자음과 자음의 접촉 유형

초성과 종성이 접촉하여 나타나는 경우로 이같은 경우는 극히 드물게 발생한다. 발생한 경우에도 그 형태가 각각 다르다. "징"의 초성과 종성의 접촉을 그 예로 들 수 있다.

⑤분리위치 지정

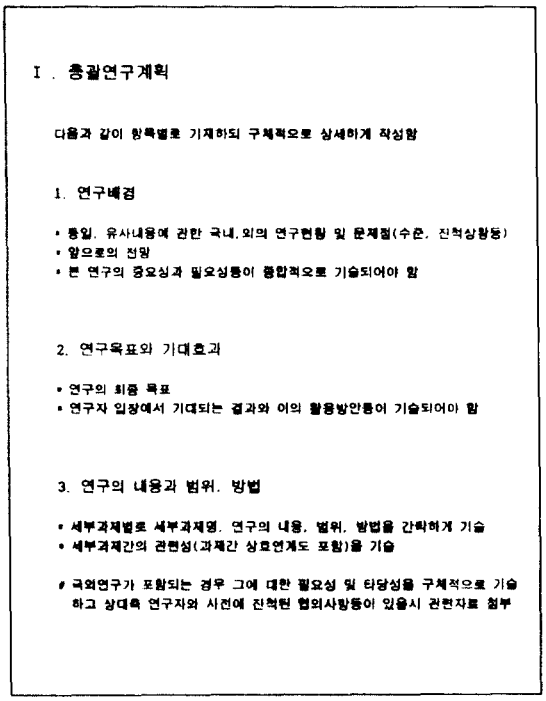
모음의 유무와 모음축에 있는 가지점과 연결된 3개의 특징점을 고려하여 분리할 특징점이 결정된다. 분리점으로 결정된 화소의 위치를 처음의 패턴에 표시하면 획의 굵기가 있기 때문에 그 굵기의 한 가운데에 놓이게 된다(그림5(c)). 이는 세선화시 굵기의 한가운데로 골격이 추출되기 때문이다. 그림5(c)에서 "c"로 표기된 부분이 분리점으로 판별된 화소이다. 분리할 방향도 각각의 접촉패턴에 따라 결정된다. 그림7의 "국"자의 예에서 초성의 "ㄱ"과 횡모음의 "ㅜ"의 접촉점의 분리방향은 점촉점의 위쪽이 된다. 정확하게 분리하기 위해서는 분리할 위치를 표시해 주어야 한다. 분리할 위치는 굵기가 작아지는 부분이 실제적으로 결정되어야 한다. 따라서 획의 굵기 값의 변화량이 급격히 일어나는 부분을 잡아주어야 한다. 접촉패턴에 따라 각각 분리할 방향이 먼저 결정되면 분리할 방향이 분리점을 중심으로 상하인지 좌우인지를 구분한다. 상하인 경우에 분리점과 연결된 수평방향의 흑화소의 수를 계산하고 좌우인 경우에 분리점과 연결된 수직방향의 흑화소 수를 계산한다. 현재 분리점의 위치를 (x, y)라 하고 수평방향으로 분리점과 연결된 흑화소의 수를 $N_h(x, y)$, 수직방향으로 연결된 흑화소의 수를 $N_v(x, y)$라 하자. 분리할 방향이 분리점을 중심으로 상하방면으로 결정되었다고 하자. 분리방향으로의 굵기에 대한 차분값은 식(1)과 같이 주어진다. 분리점으로부터 분리방향으로 좌표값이 증가한다고 하자. 차분값을 분리점으로부터 거리가 일정정도까지 측정하여 가장 큰값이 분리할 위치로 결정되어 좌표값이 식(2)와 같이 정해진다. 여기서 A는 1부터 t까지의 정수의 집합이고 t는 문자의 굵기에 따라 정해진다. 위치 값(2)에 해당하는 화소와 수평방향으로 연결된 모든 흑화소가 분리화소로 결정된다. 그림5(d)가 분리화소를 지운후의 패턴의 모습을 보여주고 있다.

$$D_h(i) = \frac{N_h(x, y+i) - N_h(x, y)}{N_h(x, y+i)} \quad (1)$$

$$(x + Max(D_h(i)_{i \in A}), y) \quad (2)$$

Ⅳ. 실험결과 및 고찰

제안한 알고리즘의 성능평가를 위하여 인쇄체 한글문서 2개를 선정하였다. 입력은 렐리시스(Relisys)사의 스콜피오(scopio)스캐너를 사용하여 A4용지 하나의 문서에 대해 각각 해상도 200 DPI(Dot Per Inch, 1728x2376)와 300 DPI(2400x3600)로 취득하였다. 그림 9는 실험에 사용된 문서를 보여주고 있다.



Ⅰ. 총괄연구계획

다음과 같이 항목별로 기재하되 구체적으로 상세하게 작성함

1. 연구배경

- 동일, 유사내용에 관한 국내.외의 연구현황 및 문제점(수준, 진척상황등)
- 앞으로의 전망
- 본 연구의 중요성과 필요성등이 종합적으로 기술되어야 함

2. 연구목표와 기대효과

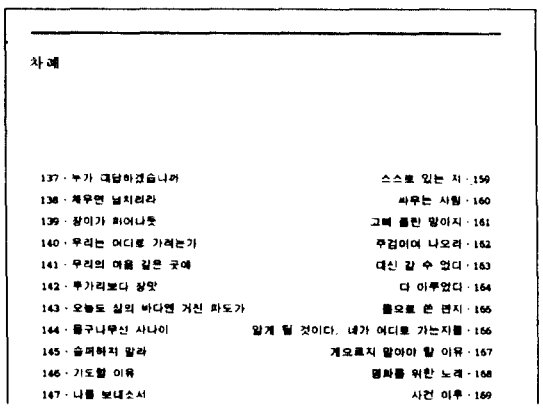
- 연구의 최종 목표
- 연구자 입장에서 기대되는 결과와 이의 활용방안등이 기술되어야 함

3. 연구의 내용과 범위, 방법

- 세부과제별로 세부과제명, 연구의 내용, 범위, 방법을 간략하게 기술
- 세부과제간의 관련성(과제간 상호연계도 포함)을 기술

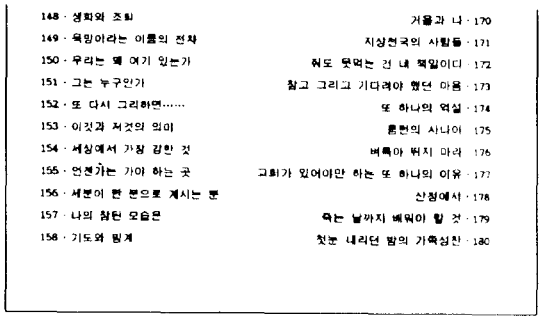
※ 국외연구가 포함되는 경우 그에 대한 필요성 및 타당성을 구체적으로 기술하고 상대측 연구자와 사전에 진척된 협의사항등이 있을시 관련자료 첨부

(a) 문서1



차례

137 · 누가 대답하겠습니까
138 · 제우면 넘치리라
139 · 장미가 피어나듯
140 · 우리는 어디로 가려는가
141 · 우리의 마음 같은 곳에
142 · 푸가리보다 장엇
143 · 오늘도 삶의 바다엔 거친 파도가
144 · 믿구나무신 사나이
145 · 슬퍼하지 말라
146 · 기도할 이유
147 · 나를 보내소서
스스로 있는 자 · 159
싸우는 사람 · 160
그뭐 좋은 말이지 · 161
주검이여 나오라 · 162
대신 갈 수 없다 · 163
다 이루었다 · 164
돌으로 쓴 편지 · 165
알게 될 것이다, 내가 어디로 가는지를 · 166
계오르지 말아야 할 이유 · 167
평화를 위한 노래 · 168
사건 이후 · 169



(b) 문서2

그림 9. 실험에 사용된 문서

문서1은 연구보고서 작성요령에 관한 내용을 담고 있고 문서2는 어느 책의 차례를 담고있는 내용을 스캐닝한 것이다. 표1은 기존의 방법 즉 유손실인 경우에는 PMS[5], 무손실인 경우에는 SPM[7] 무손실모드, 표2는 제안한 방법으로 처리하였을 경우 추출된 패턴의 수와 매칭비율을 나타낸 것이다.

표 1. 기존 알고리즘[5, 7]에 의한 패턴매칭 비율
Table 1. The match ratio by existing algorithm[5, 7]

문서	해상도 (DPI)	추출된 패턴 수	매칭비율	
			무손실[12]	유손실[9]
문서1	200	629	75.7%	71.5%
	300	642	76.6%	74.9%
문서2	200	1012	86.0%	82.5%
	300	1015	86.2%	84.7%

표 2. 제안한 알고리즘에 의한 패턴매칭 비율
Table 2. The match ratio by proposed algorithm

문서	해상도 (DPI)	추출된 패턴 수	매칭비율	
			무손실	유손실
문서1	200	766	81.3%	77.2%
	300	774	79.8%	77.9%
문서2	200	1103	88.1%	83.8%
	300	1105	87.9%	86.4%

전반적으로 유손실보다 무손실이 매칭비율이 높은 것은 무손실 모드인 경우에는 부호화 할 패턴이 매칭된 패턴일지라도 라이브러리에 새롭게 추가되기 때문에 라이브러리 크기가 유손실의 그것 보다 크기 때문이다(그림4 참조). 동일문서인 경우에 해상도가 높은 문서에서 추출된 패턴 수가 많은 것은 입력과정에서 해상도가 낮은 경우에는 패턴의 크기가 작아서 잡음으로 처리되었던 일부가 해상도가 높아짐에 따라 패턴으로 처리되었기 때문인 것으로 생각된다. 제안한 알고리즘의 추출된 총 패턴의 수가 기존의 방법보다 많은 것은 자소분리를 함으로서 패턴이 분리되어 생긴 결과이다. 제안한 알고리즘의 매칭율이 기존의 패턴매칭에 의한 방법에 비해 1.3%~5.7% 정도 향상된 것을 알 수 있다. 자소분리로 인한 패턴의 수가 증가하여 위치정보라든가 패턴의 크기등에 관한 정보량은 기존의 알고리즘에 비해 증가하였고 비트맵에 관한 정보량은 감소된 경향을 나타내고 있다. 표3는 그동안 발표되어진 2진영상 압축 알고리즘과 제안한 알고리즘의 컴퓨터 모의실험한 결과 압축된 영상의 비트(bit)수를 보여주고 있다. MR 알고리즘은 K=4로 실험하였다. JBIG은 시퀀셜모드로 실험한 결과이다. MR, MMR, JBIG, SPM(무손실), 제안한 무손실 모드, PMS(유손실), 제안한 유손실 모드 순서로 성능이 향상되는 경향을 보여주고 있다. 모든 알고리즘이 해상도가 높을수록 압축율은 증가하는 경향을 나타내었다.

기존의 패턴매칭에 의한 방법[9], [12]과 제안한 알고리즘은 공히 새로운 패턴 여부를 판단하기 위한 임계치로써 패턴의 전체크기에 대한 오차화소의 비를 21%로 사용하였다. MMR 알고리즘의 압축율은 MR의 그것에 비해 1.5-1.8배 향상된 결과를 나타내었다. JBIG은 MMR에 비해 압축율이 23-25%정도 향상되었다. 기존의 패턴매칭에 의한 방법[12]의 무손실모드는 JBIG에 비해 13-38% 정도의 압축율 향상이 이루어 졌다. 패턴 매칭기법에 의한 유손실모드는 무손실에 비해 대략 2배 정도 우수한 경향을 띠고 있다. 이는 일정정도의 화소의 손실을 가져온 댓가이다. 제안한 알고리즘의 유손실은 무손실에 비해 약 1.8~2.4배 정도 압축율이 우수했다. 제안한 알고리즘의 무손실인 경우는 JBIG과 비교했을때 압축율이 17~50%정도 향상되었고 유손실인 경우에는 JBIG에 비해 2.1배에서 최고 3.3배까지 향상되었다. 제안한 알고리즘은 무손실인 경우 기존의 SPM의 무손실모드에 비해 알고리즘에 비해 압축율이 1.3%에서 3.0%까지 향상되었고 유손실인 경우에는 기존의 PMS보다 3.4%에서 9.1%까지 향상되었다. 자소분리과정에서 일부 부정확한 분리가 되는 경우가 있었다. 예를 들면 그외에

표 3. 결과(bits)
Table 3. Results(bits)

문서	해상도 (DPI)	MR	MMR	JBIG	SPM[12] (무손실)	PMS[9] (유손실)	제안안	
							무손실	유손실
문서1	200	137415	78601	62741	54201	32563	53517	29835
	300	213079	116795	94909	76025	40465	74593	37226
문서2	200	186915	118888	94764	76574	38031	75513	36599
	300	289007	175304	142417	103059	45511	101517	42214

"답"자 같은 경우는 "ㅏ"와 "ㅂ"이 자연스럽게 붙어 있어 접촉점에서 특징점이 추출되지 않아 부정확하게 분리되었다. 이러한 부정확한 불리는 성능향상을 저해하는 요소이나 그 비중이 작아 전체시스템에 미치는 영향은 미약하였다. 자소분리과정으로 인하여 계산량은 약간 늘어 났으나 전체 계산량에 비하면 무시할 정도이다. 특히 유손실인 경우에는 라이브러리 크기가 기존의 방법에 비해 상대적으로 작아짐으로써 라이브러리 탐색하는 계산량이 감소하는 경향도 나타내었다. 유손실모드에서 화소단위의 비트에러율은 0.3%-0.5%정도로 기존의 알고리즘[9]과 제안한 알고리즘이 거의 유사하게 나왔다. 잘못 매칭되는 경향은 2개의 알고리즘이 서로 비슷하게 나타났다. 한글은 영문자에 비해 상대적으로 유사문자가 많기때문에 유손실모드인 경우에 오매칭 개선을 위한 연구가 앞으로 필요할 것으로 생각된다.

다른 앞으로의 연구 과제로 도면과 같은 문자가 아닌 특수한 응용문서를 대상으로 상관성을 개발하는 알고리즘에 대한 연구도 필요하리라 생각된다. 그리고 좀 더 유연하고 정확한 분리 알고리즘에 대한 연구가 지속적으로 요구된다. 본 논문에서는 고딕체 문서를 대상으로 실험하였다. 따라서 분리 알고리즘에 사용된 여러가지 임계값(예를 들면 모음의 유무 판별시 끝점의 탐색영역등)이 각 문체에 따라 적응적으로 처리되는 일반화된 분리 알고리즘에 대한 연구가 필요할 것으로 생각된다.

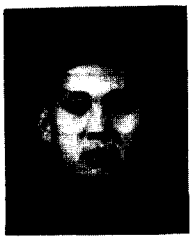
V. 결 론

패턴매칭과 산술부호화를 근간으로 하는 알고리즘이 비교적 최근에 개발된 2진영상을 대상으로 하는 우수한 압축 알고리즘이다. 패턴매칭의 관점에서 볼 때 한글은 자소의 불규칙한 접촉으로 인한 다양한 패턴의 발생으로 본래의 취지를 떨어뜨릴 수 있다. 패턴간의 상관성을 효율적으로 개발하기 위해 자소의 접촉점을 추출하여 분리함으로 자소간의 매칭이 이루어지도록 유도하였다. 자소분리를 하지않았을 경우에 비해 매칭비율이 1.3-5.7%까지 증가하여 기존의 알고리즘[9, 12]에 비해 무손실인 경우에는 최대 1.3-3.0%, 유손실인 경우에는 3.4-9.1% 압축율이 향상되었다.

참 고 문 헌

1. New work item proposal: JPEG-2000 image coding system. ISO/IEC JTC/SC29 /WG1
2. R. Hunter and A. H. Robinson, "International Digital Facimile Coding Standards," Proc. OF The IEEE, Vol. 68, No. 7, pp. 854-867, July. 1980.
3. W. B. Pennebaker and J. L. Mitchell, "JPEG: still image data compression standard" Van Nostrand Reinhold 1993.
4. R. N. Ascher and G. Nagy, "A means for achieving a high degree of compaction on scan-digitized printed text," IEEE Trans Comput., vol. C-23, no. 11, pp. 1174-1179, Nov. 1974.
5. O. Johnsen, J. Segen and G. L. Cash, "Coding of Two-Level Pictures by Pattern Matching and Substitution." Bell System Technical Journal Vol. 62, No. 8, Oct 1983.
6. I. H. Witten, T. C. Bell, H. Emberson, S. Inglis and A. Moffat, "Textual Image Compression: Two-Stage Lossy/Lossless Encoding of Textual Images" Proc. of The IEEE, VOL. 82, NO. 6, pp. 878-888. June. 1994.
7. P. G. Howard, "Lossless and Lossy Compression of Text Images by Soft Pattern Matching," A pro-

posal submitted for JTC 1. 29. 10 [JTC 1/SC29/WG1]
8. 남궁재찬 "Index-Window 알고리즘에 의한 한글 pattern의 부분분리와 인식에 관한 연구," 인하대학교 박사학위 논문, 1982.
9. 최필웅, 이기영, 구하성, 고 형화, "접촉점에서의 부분그래프 패턴에 의한 필기체 한글의 자소분리에 관한 연구," 대한전자공학회 논문지, 제30권, B편 제4호, pp. 254-273, 1993년 4월.
10. Y.K. Chu and C. Y. Suen, "An alternate smoothing and stripping algorithm for thinning digital binary patterns," Signal Processing, vol. 11, no. 3, pp. 207-227, 1986.



김 영 태(Kim Young Tae) 정회원
1991년 2월 : 광운대학교 전자통신공학과 공학사
1993년 2월 : 광운대학교 전자통신공학과 공학석사
1993년 3월~현재 : 동대학원 박사과정 재학중
※주관심분야 : 디지털 영상처리, 데이터 압축, 패턴인식

고 형 화(Ko Hyung Hwa) 정회원
현재 : 광운대학교 전자공학부 교수
한국통신학회 논문지 제21권 제10호 참조