만주문자 인식을 위한 전처리 방법에 관한 연구
A Study on the Preprocessing for Manchu-Character Recognition

최민석*, 이충호*
Minseok-choi*, Choong-ho Lee*

요 약

만주문자로 기록된 문헌의 디지털화에 대한 연구는 아직 초기 단계이다. 본 논문은 만주문자의 인식을 위한 전처리 방법을 제안한다. 만주문자의 전처리 단계는 세선화와 문자단위 분리가 중요하다. 본 논문에서는 기존 세선화 방법인 Hilditch 세선화 알고리듬을 개선하여 만주문자의 세선화 오류를 보완하고 각 문자단위를 좌우측으로 분류하지 않고 문자의 삐침이 존재하는 위치점 사이의 중심점을 이용하여 분리하여 내는 실제적인 방법을 제안하고 있다. 실험을 통하여 만주문자로 이루어진 단어의 세선화와 문자단위 분류에 적용하여 그 유효성을 보여주고 있다.

ABSTRACT

Research for Manchu character digitalization is at an early stage. This paper proposes a preprocessing algorithm for Manchu character recognition. This algorithm improves the existing Hilditch thinning algorithm so that it corrects thinning error for Manchu characters. The existing algorithm separates the characters into the left-hand side and right-hand side, while our alogorithm uses the central point between the points that strokes exist when it classifies each of characters. The experimentation results show that this method is valid for thinning and classification of Manchu characters.

Keywords : Manchu character recognition, Hilditch thinning algorithm, character recognition

Ⅰ. 서 론

만주문자로 기록된 문헌이 중국과 한국 등에 약 200만권에서 300만권에 이르는 방대한 양이 존재하는 것으로 알려져 있으나 정치적, 문화적, 기술적 이유 때문에 디지털화에 대한 연구가 미미한 상태이다. 정치적으로는 만주족이 중국에 속해 있어 만주어를 사용하던 만주족이 더 이상 만주어를 더 이상 사용하지 않고 있어서 만주고유문자 해독 인력에 별도의 시간과 비용이 소요되기 때문이다. 더욱이 만주고유문자는 그 특성상 디지털화나 자동인식이 매우 어려운 글자에 속한다. 왜냐하면 만주문자는 어두, 어중, 어말에서 형태가 현격히 달라지는 경우가 많으며, 같은 글자를 중복하여 다른 문자로 사용하는 경우가 많기 때문이다. 따라서 이를 해독하여 전사(Romanization)하는 인력양성에 많은 시간과 노력이 필요하다. [1-4]

이런 상황에서 최근 만주문자를 디지털신호처리기술을 이용하여 자동 인식하고자 하는 연구[5,6]가 몇몇 보고 되고 있다. 이 방법은 만주문자의 중심획을 따라서 문자를 양쪽으로 분류하고 그 양쪽으로 분류된 획과 점등을 조합하여 문자를 인식하는 방법이다. 하지만, 이런 연구들은 아직 초기단계로서 다른 연구자들에 의하여 그 유효성이 충분히 입증되었다고는 볼 수 없다. 이런 방법은 기존의 문자인식과는 현격히 다른 방법으로서 이런 방법이 최선을 가져 오는지에 대한 비교연구가 이루어지기 위해서는 다른 방법들이 병행하여 연구되어야 하기 때문이다.

한편 만주문자를 전처리 하기 위해 사용될 수 있는 전처 알고리듬 중에서 많이 알려진 것으로는 Hilditch 알고리듬[7-8]과 Zhang-Suen 알고리듬[9-10]이 있으며 이 외의 어떤 알고리듬이든 모두 각기 대상패턴에 대한 장단점을 가지고 있다. 따라서 현재는 이들 방법 또는 다른 방법을 조합하여 특정 대상 패턴에 사용하려는 연구가 이루어지고 있다[8,10]. 이들 알고리듬은 적용 대상패턴에 따라서 대상화소를 삭제하는 조건을 줄 때에 타협점을 찾아야 한다.

본 논문에서는 기존의 만주문자를 양쪽으로 분리방법을 사용하지 않는 다른 방법을 제안하는 동시에, 만주문자의 자동인식을 위한 첫 단계로서 만주어의 전처리과정인 세선화와 만주문자 단위를 단어로부터 분리하는 방법을 제안한다. 기본적인 전처리 과정은 Hilditch 세선화 방법을 수정하여 적용하는 방법을 사용한다. 본 논문의 구성은 다음과 같다. 제 II 장에서는 먼저 배경지식이 되는 만주문자의 특징 및 관련연구에 대하여 기술한다. 제 III장에서는 제안하는 알고리듬을 설명한다. 제 IV장에서는 실험결과에 대하

* 한발대학교 정보통신공학과
투고 일자 : 2013. 1. 9 수정완료일자 : 2013. 4. 15
게재확정일자 : 2013. 4. 30
※본 논문은 2010년도 한발대학교 교내학술연구비의 지원으로 수행되었습니다.

여 간략하게 설명하고 제 V장에서 현 단계에서의 결론을 맺는다.

II. 만주문자의 특징 및 관련연구

만주문자는 국내에서 거의 연구되고 있지 않아 이에 대한 설명이 필요할 것으로 생각되므로, 먼저 만주문자의 특징에 대하여 설명하고 그 다음 만주문자 인식을 위한 기존 연구에 대하여 기술하기로 한다.

2.1 만주문자의 특징

만주문자는 그림 1과 같이 그 문자단위가 놓이는 위치에 따라 같은 문자가 어두, 어중, 어말에 따라 형태가 달라지거나 중복되어 사용된다. 즉, 하나의 단어는 여러 개의 문자가 중심선을 기준으로 이어진 형태로 구성되고 있으며, 같은 문자라 할지라도 단어의 첫 글자에 사용하는 것, 중간에 사용하는 것, 마지막 글자에 사용하는 것, 그리고 한글자만 단독으로 표기할 때 쓰는 문자가 각기 다른 형태를 갖고 있다. 이것은 한국의 한글이나 일본의 히라가나, 카타카나의 각 문자가 문장의 각 위치에서 변형되는 정도가 획의 삐침이나 크기의 변화가 약간 존재하는 것과는 달리, 그 모양이 사람이 같은 문자로 인식하기 어려울 정도로 현격하게 달라진다는 점에서 다르다. 한편 만주문자에서 한 글자가 다른 발음을 위하여 중복되어 사용되는 경우는 한글에서 '이응'이 초성으로 쓰이는 경우 묵음이며 종성에서만 발음이 되는 것, 그리고 기역, 쌍기역, 디귿, 쌍디귿, 시옷, 쌍시옷, 히읗 등이 초성과 종성에서 발음이 다른 것과 개념상 비슷하다. 하지만 그 모양 면에서는 다른 위치에서 완전히 달라지는 경우도 있고 앞뒤의 글자에 따라 일정 발음을 나타내는 글자가 완전히 다른 경우가 있다.

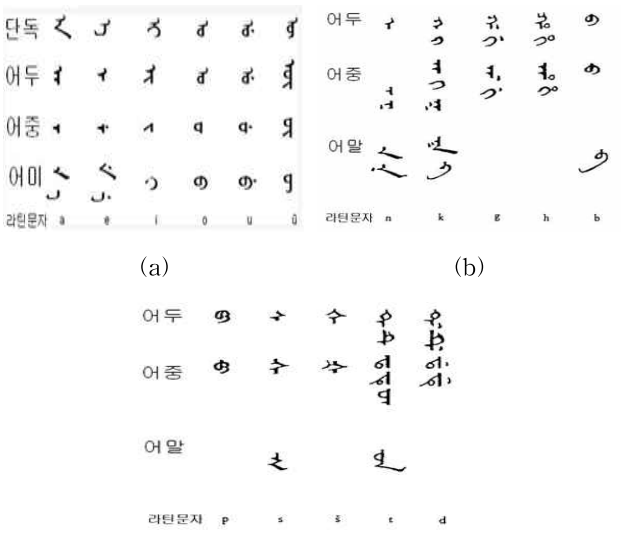

그림 1. 만주문자의 어두, 어중, 어말에 따른 형태의 변화 로마자 표기 일례: (a) 모음; (b)와(c) 자음

Fig. 1. An example for Romanization of Manchu characters and the isolated form, initial form, medial form and final form: (a) vowels; (b) and (c) consonants.

만주 고유문자를 기록하는 방법은 그림 2와 같이 위에서 아래쪽으로 세로로 이어서 쓰도록 되어 있으며, 각 세로줄은 왼쪽에서 오른쪽으로 순서대로 써 나가도록 되어 있다.

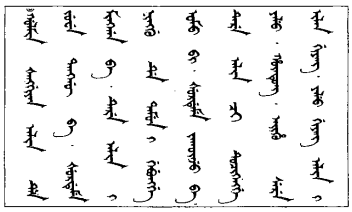

그림 2. 만주문자로 기록된 문헌의 예

Fig. 2. Example of documents written in Manchu characters.

만주어로 기록된 문헌은 인쇄체와 필기체가 있으나 대부분이 필기체로 기록되어 그 인식이 매우 어렵다.

2.2 기존의 방법

기존의 방법[5,6]에서는 Hilditch 방법으로 세선화 한 후 중심선에서 좌우로 임계치를 주어 중심에서 임계치내에 있는 화소들을 삭제하는 방법을 취하였다. 또한 문자의 분리는 중심획을 중심으로 우방향 연결된 부분, 우방향 절단된 부분, 좌방향 연결된 부분, 좌방향 절단된 부분으로 나누어 인식할 수 있도록 처리하였다.

Hilditch의 세선화 알고리듬[7]은 다음과 같다. 그림 3은 세선화 알고리듬 적용을 설명하기 위한 중심화소 $P(i,j)$와 이웃화소를 보여 준다. 가로방향은 j, 세로방향은 i로 나타내었다.

$P_{i-1,j-1}$	$P_{i-1,j}$	$P_{i-1,j+1}$
$P_{i,j-1}$	$P_{i,j}$	$P_{i,j+1}$
$P_{i+1,j-1}$	$P_{i+1,j}$	$P_{i+1,j+1}$

그림 3. 세선화를 적용할 중심화소와 이웃화소들

Fig. 3. Central piexel and neighboring pixels for thinning.

아래의 조건들을 모두 만족하면 i,j의 흑색화소를 삭제하는 것이다. 여기서 이웃한 흑색 화소의 개수를 $B(i,j)$라고 하고 세선화 중심화소의 이웃화소를 한 바퀴 회전하며 검색할 때 흑색화소에서 백색화소로 바뀌는 횟수 (혹은 백색 화소에서 흑색화소로 바뀌는 횟수)를 $A(i,j)$로 나타낸다.

조건 1) 이웃한 검정화소가 2~6개 사이일 때 $(2 \leqq \mathrm{B}(i,j) \leqq 6)$.
조건 2) 이웃화소를 회전할 때 흑색에서 백색으로 변화되거나 혹은 백색에서 흑색으로 변화되는 횟수가 한 번인 경우. 즉 $\mathrm{A}(i,j)=1$인 경우.
조건 3) 기준이 되는 중심화소의 좌측, 상단, 우측에 위치하는 화소 중 하나가 백색화소 일 때.
즉, $\mathrm{P}(i,j-1)+\mathrm{P}(i-1,j)+\mathrm{P}(i,j+1)\neq 0$일 때.
조건 4) 기준이 되는 중심 화소의 상단, 우측, 하단 화소 중 하나가 백색화소 일 때.
즉, $\mathrm{P}(i-1,j)+\mathrm{P}(i,j+1)+\mathrm{P}(i+1,j)\neq 0$ 일 때.
이 방법을 만주문자에 대하여 적용한 결과는 그림 4와 같다.

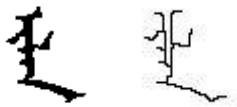

그림 4. Hilditch 세선화 알고리듬을 적용한 결과
Fig. 4. Result which Hiditch thinning algorithm is applied.

그림 4와 같이 중심선이 2개로 분리되는 문제점이 있음을 알 수 있다. 이런 현상이 일어나는 이유는 Hilditch 세선화 방법을 적용하였을 때 발생하는 문제점으로서 그림 5의 (a), (b)와 같은 경우에 해당되어 세선화가 올바르게 적용되지 않기 때문이다. 그림 5의 (a)는 중심화소가 백색이어서 세선화 처리가 되지 않으며, 그림 5의 (b)는 이웃하는 흑색화소의 개수 B(i,j)가 7이기 때문에 이웃하는 화소가 2개에서 6개일 때만 처리하도록 되어 있는 조건 1의 경우에 해당되어 마찬가지로 처리되지 않는다. 그림 5의 (c)의 경우에는 조건 1과 조건 2를 만족하지만 조건 3과 조건 4가 만족되지 않아 역시 세선화 처리가 되지 않는다.

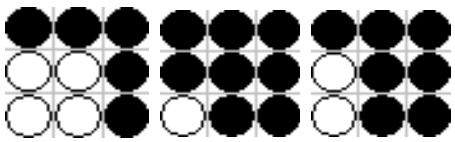

그림 5. Hilditch 세선화 알고리듬이 적용되지 않는 패턴: 좌측에서 우측으로 (a), (b), (c)
Fig. 5. The pattern that Hilditch alogorithm cannot be applied: from left to right (a),(b) and (c).

또한 3), 4)번 조건에 결점이 있는데 상단이나 우측 화소 중에서는 하나만 백색이어도 된다. 그러나 둘중 어디에도 백색이 없다면 하단화소와 좌측화소가 둘 다 백색이어야만 중심화소를 지운다. 따라서 그림 6같은 경우 중심화소가 백색으로 변화되는 세선화 오류가 발생한다.

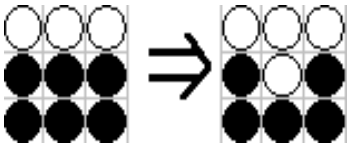

그림 6. 조건 3), 4)에 의한 세선화 오류 발생의 경우
Fig. 6. A case that thinning error is occurred by condition 3) and 4).

단위 문자의 분리 방법은 그림 7와 같이 좌측 연결부분, 좌측 비연결 부분, 우측 연결부분, 우측 비연결 부분으로 분류하여 후에 조합하는 방식을 취하고 있다.

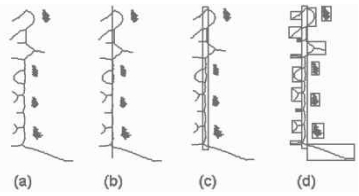

그림 7. 기존의 문자단위 분류 방법: (a) 전처리과정을 거친 상태 (b) 중심선 구하기, (c) 중심선 영역 확장, (d) 분류
Fig. 7. Existing classifying method by character units: (a) after preprocessing (b) obtaining cental line (c) extending central-line region (d) classification.

Ⅲ. 제안하는 알고리듬

상기 기술한 문제점을 해결하기 위하여 본 논문에서는 Hilditch 알고리듬의 조건 3), 4)를 수정하여 적용한다. 중심화소의 상단화소가 그림 6과 같이 잘못 변화되는 문제점을 고려하여, 중심선이 둘로 갈라지도록 하지 않게 하는 방법을 다음과 같이 제시한다. 제 Ⅱ장에서 전술한 Hilditch 알고리듬에서 조건 3) 조건 4)를 각각 아래와 같이 수정하였다.

조건 3) 기준이 되는 중심화소의 상단, 우측, 하단 중 하나가 백색화소 일 때.
즉, $\mathrm{P}(i-1,j)+\mathrm{P}(i,j+1)+\mathrm{P}(i+1,j)\neq 0$일 때.
조건 4) 기준이 되는 중심 화소의 우측, 하단, 좌측 화소 중 하나가 백색화소 일 때.
즉, $\mathrm{P}(i,j+1)+\mathrm{P}(i+1,j)+\mathrm{P}(i,j-1)\neq 0$ 일 때.
중심화소를 백색으로 바꾼다.(그림 8 참조.)

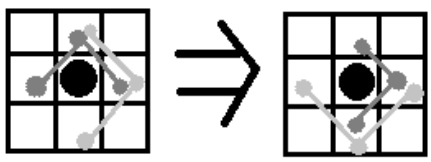

그림 8. 수정된 세선화 알고리듬의 조건: 좌측부터 (a), (b). 진한 회색은 조건 3), 옅은 회색은 조건 4)이다.
Fig. 8. The modified thinning condition: from left to right (a), (b) dark gray denotes condition 3) and light gray denotes condition 4.

다음으로 문자단위를 구분하기 위한 방법에 대하여 기술한다. 기존 방법과 같이 좌우의 획을 분리하지 않고 좌우의 획이 뻗어나간 점들의 사이에 존재하는 중심점에 경계선을 정해주는 방법을 택하였다. 세선화과정에서

중심선이 약간 틀어질 수 있는데, 이를 고려하여 오차범위를 중심선 좌측 화소부터 우측 2화소까지 잡는다. 그 안에서 같은 세로줄에 흑색화소가 1~3개이고, 같은 줄에서 해당영역을 벗어난 곳에는 흑색화소가 없을 때 삐침이 아닌 영역으로 판단한다. 그 영역의 중간 위치를 계산해서 구하면 삐침사이의 경계선이 되는 것이다.

위에서 기술한 방법에 따라, 참고문헌 [4]에서 임의로 취한 단어 중 한 개를 제 그림 4의 왼쪽 그림에 적용하면 그림 9와 같으며, 양호한 결과임을 알 수 있다.

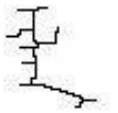

그림 9. 제안된 방법으로 그림 6을 세선화한 결과
Fig. 9. A result after thinning using the proposed method.

그림 9와 같이 세선화를 행 한 후 문자단위를 구분하기 위한 경계선을 결정하는 방법에 대하여 기술한다. 다음과 같은 사항들을 고려하였다.

1) 왼쪽 삐침사이의 중간 위치에 적용한다.

2) 세선화 하는 과정에서 중심선 화소가 1~2 화소 정도 어긋나는 경우가 있다.

3) 어두를 구성하는 삐침사이의 중간 위치는 적용 대상에서 제외한다.

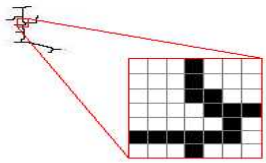

그림 10. 중심선이 오른쪽으로 2화소 정도 어긋나 있는 경우
Fig. 10. When the central line is dislocated by two pixels toward left hand.

이러한 점들을 염두에 두고 다음과 같은 순서대로 경계선을 정한다.

1) 한 행에서 중심선을 기준으로 -3부터(좌) +3까지(우) 흑색화소의 개수를 더한다.

2) 흑색화소가 1~3개 이면서 중심선을 기준으로 -4와(좌) +4의(우) 화소가 백색이면 삐침이 없는 위치로 판단한다.

3) 삐침이 없는 행부터 삐침이 있는 행이 나오기 전까지 행의 개수를 누적한다.

4) 누적한 값을 이용하여 중간 값을 구한다.

중간 값이 가리키는 위치가 경계선을 설정하는 위치가 된다.

5) 마지막으로 어두를 구성하는 부분에 적용된 경계선만 무효화시킨다.

위의 방법을 따라 경계선을 적용한 결과는 그림 11과 같다.

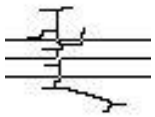

그림 11. 그림 9에 경계선을 적용한 결과
Fig. 11. Result for dividing lines are drawn.

Ⅳ. 실험 결과

본 논문에서는 제안된 알고리듬으로 세선화를 행하고, 만주문자의 경계를 긋는 데까지가 실험의 범위이다. 실험에 사용된 만주어 단어는 참고문헌 [4]에서 임의로 취한 것을 사용하여 수행하였다. 우리가 아는 한 아직까지는 공인된 만주문자 데이터베이스가 존재하지 않아, 일단 본 논문의 목적상 제안된 방법의 유효성을 검증할 수 있는 범위로 실험을 제한하였다.

제 Ⅲ장에서 기술한 방법을 적용하여 그림 9와 같이 세선화된 결과를 얻은 다음, 그림 11과 같은 결과를 얻기 위하여 경계선을 결정하는 방법을 적용하였다. 이 같은 과정을 10개의 단어에 적용하여 본 결과는 그림 12와 같다.

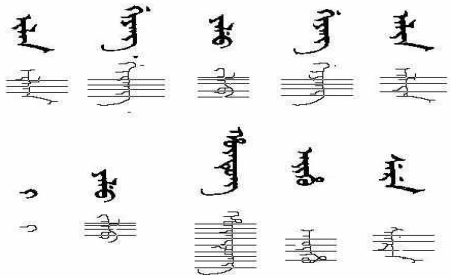

그림 12. 만주어 단어 10가지의 세선화, 경계선설정을 한 결과
Fig. 12. Thinning and drawing of dividing lines for 10 Manchu vocabulary.

경계선을 설정한 결과, 육안으로도 어떤 글자인지 판별하기 비교적 용이하여졌음을 알 수 있다.

다만 예외적으로 의도한대로 경계선 설정이 되지 않는 것도 있다. 가령 모음 중에 어두에 사용되는 'o'나 'u'에 해당하는 만주문자는 위의 방법대로 인식시키면 삐침이 3번 정도 인식이 된다. 그렇게 되면 2번째와 3번째의 삐침 사이의 경계선이 남게 된다. 이런 식으로 원래 하나인 문자가 둘로 나뉘어 인식 되는 경우가 발생할 수가 있다. 이것은 만주어 어휘를 기반으로 한 맞춤법 검사로 정정이 필요하다.

가령 아래와 같은 만주어의 음운론적 특징들을 적용하는 방법이 있다.

1) w 뒤에는 a, e 이외의 모음이 오지 않는다.
2) 차용어를 제외하고 t, d 뒤에 i가 오지 않는다.
3) y 뒤에 모음 i가 오지 않는다.
4) ū는 보통 k, g, h 뒤에만 올 수 있다.
5) r은 어두에 오지 않는다.
6) ng는 음절 초에 오지 않는다. 또 어중의 ng는 k, g 앞에만 나타난다.
7) 음절 말에 올 수 있는 자음은 b, m, t, n, r, l, s, k, ng이다.
8) 어말에 올 수 있는 자음은 n뿐이다. (외래어 제외)
9) 보통 음절 초 또는 음절 말에 자음이 연속되지 않는다. (음절 말 자음과 음절 초 자음은 연속될 수 있음)

또한 반대로 나누어져야 할 2개의 문자가 분리 되지 않는 경우가 있다. 이러한 경우는 2개의 조합을 하나의 새로운 문자처럼 인식하도록 정해놓는 방법도 생각할 수 있다.

본 논문의 실험에서는 10개의 만주문자로 이루어진 단어를 사용하여 제안한 세선화 알고리듬과 문자단위 분리방법을 적용한 결과, 세선화는 10개 모두 성공하여 100% 성공하였고, 문자단위 분류는 9개가 성공하여 90%의 성공률을 보였다.

V. 결 론

본 논문에서는 만주문자 인식을 위한 전처리 단계로서 Hilditch 알고리듬을 수정하여 적용하였고, 만주문자를 문자단위로 분류하는 새로운 방법을 제안하였다. 실험에서는 10개의 만주문자로 이루어진 단어를 사용하여 제안한 세선화 알고리듬과 문자단위 분리방법을 적용한 결과, 세선화는 10개 모두 성공하여 100% 성공하였고, 문자단위 분류는 9개가 성공하여 90%의 성공률을 보였다. 주어진 실험데이터 상에서는 제안된 방법의 유효성을 확인하였다. 향후 과제는 많은 데이터를 통하여 제안된 방법을 개선하고, 궁극적으로는 만주문자를 자동 인식하는 문제가 남아 있다.

참 고 문 헌

[1] 莊吉發編譯, 御門聽政 -滿語對話選萃-, 文史哲出版社印行, 1999.
[2] 김득황, 기초만한사전, 대지문화사, 1997.
[3] 羽田亨, 滿和辭典, 學海出版社, 1998.(再版)
[4] 河內良弘, 滿洲語文語辭典, 京都大學出版社, 1996.
[5] G. Y. Zhang, J. J. Li and A. X. Wang, "A new recognition method for the handwritten Manchu chracter unit," Proc. of the 5th Int'l Conf. on Machine Learning and Cybermetics, pp. 3339-3344, Aug. 2006.
[6] G. Y. Zhang, J. J. Li, "An offline recognition method of handwritten primitive Manchu characters based on strokes," Proc. of the 9th Int'l Workshop on Frontiers in Handwriting Recognition, IEEE Computer Society, 2004.
[7] C. J. Hilditch, Linear Skeletons from Square Cupboards, in Machine Intelligence IV (B. Meltzer and D. Mitchie eds), University Press, Edinburgh, 1969.
[8] J. Yu and Y. Li, "Improved Hilditch thinning algorithms for text image," Int'l Conf. on E-Learning, E-Business, Enterprise Information Systems and E-Government, pp. 76-79, 2009.
[9] T. Y. Zhang, C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," Comm. ACM, Vol 27. pp. 236-239, 1984.
[10] W. Chen and L. Sui, "Improved Zhang-Suen thinning Algorithm in bnary line drawing applications," Int'l Conf. on Systems and Informatics, pp. 1947-1950, 2012.

최 민 석(Min-seok Choi)

2011년 8월 한밭대학교 정보통신컴퓨터공학부 정보통신공학전공(학사)
2012년 3월 ~ 현재 한밭대학교대학원 정보통신전문대학원 정보통신공학과 재학 중
※주관심분야 : 패턴인식, 디지털신호처리

이 충 호(Choong-ho Lee)
正會員

1985년 2월 연세대학교 전자공학과(학사)
1987년 2월 연세대학교대학원 전자공학과 (석사)
1998년 3월 토호쿠대학대학원 정보과학연구과 시스템정보과학전공 (공학박사)
1987년 2월~ 2000년 2월 KT 멀티미디어연구소 전임연구원
2000년 2월 ~ 현재 한밭대학교 정보통신공학과 교수
※주관심분야 : 패턴인식, 디지털신호처리, 응용소프트웨어