서식 문서의 선과 접촉된 숫자열 복원에 관한 연구

(Restoration of Numeral Strings Touched with Lines in Various Form Documents)

이 창 현 † 최 영 우 †† 김 경 환 ††† 이 일 병 ††††

(Changhyun Lee) (Yeongwoo Choi) (Gyeonghwan Kim) (Yillbyung Lee)

요 약 본 논문에서는 서식 문서의 선과 숫자의 획이 접촉된 경우 숫자의 획을 접촉되기 전 상태의 원 이미지로 복원하는 방법을 제안한다. 제안하는 방법은 서식 문서에서 추출한 숫자열을 대상으로 열 단위로 복원한다. 과정은 우선 숫자열과 접촉된 선의 위치를 찾아내고, 선을 추적하면서 접촉으로 판정되는 영역을 유형별로 분류하여, 각 유형에 적합한 획 복원 방법을 제안한다. 또한 선에 숫자의 획이 완전히 포함된 경우의 복원 방법도 제안하여 현장에서의 서식 처리 과정에서 발생하는 문제점을 해결하고자 하였다. 제안하는 방법을 평가하기 위해서 은행 입출금전표, 신용카드 매출전표 및 NIST 필기 숫자열 데이타베이스 이미지를 사용하였다.

Abstract This paper presents a method for restoring numeral strings when lines are touched with numeral strokes in various form documents. The proposed method restores the numeral string without segmenting them into individual numerals. It first finds the location of the touching lines in the extracted field image, and traces the lines and classifies the touching area into several predetermined types. Then each area is restored according to the classified types of touching. In this paper we also try to restore strokes when the strokes are totally included in the form lines to solve the difficulties in real forms processing applications. For the evaluation we have applied the method to the various handwritten and machine-printed numeral strings from bank slips, credit card sales slips, and NIST database images.

1. 서 론

서식 문서는 대부분 수평선과 수직선으로 이루어진 사각형 모양의 박스 안에 문자열이 기입되지만, 서식의 선과 문자열이 접촉되는 경우가 빈번하게 발생된다. 그림 1은 서식 문서에 기입된 두 종류의 숫자열 이미지로서 서식의 선에 숫자 획의 일부가 접촉되거나 획 전체가 포함된 경우를 보여준다. 서식 자동 입력을 위해서는 이와 같이 서식의 선에 접촉되거나 포함된 숫자의 획을 복원하는 과정이 필요하다.

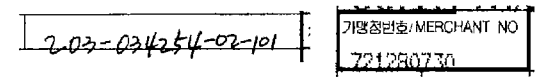


그림 1 서식의 선에 숫자 획이 접촉되거나 포함된 예

위의 문제를 처리하기 위한 간단한 방법은 다음과 같다. 우선 선의 일부를 포함시킨 상태에서 문자 이미지를 추출한 연구와[1,2], 문자열에 접촉된 선을 파악하고 선을 무조건 제거함으로써 문자 이미지가 훼손된 상태로 추출되는 연구[3,4] 결과가 있었다. 선의 일부가 문자 이미지에 포함된 상태로 추출된 경우나 문자 획의 일부

† 비 회 원 : 삼성SDS 연구원
aihyun64@samsung.co.kr
†† 정 회 원 : 숙명여자대학교 정보과학부 교수
ywchoi@sookmyung.ac.kr
††† 정 회 원 : 서강대학교 전자공학과 교수
gkim@ccs.sogang.ac.kr
†††† 종신회원 : 연세대학교 컴퓨터과학과 교수
yblee@csai.yonsei.ac.kr
논문접수 : 2000년 2월 17일
심사완료 : 2001년 4월 2일

가 소실되어 추출된 경우 모두 신뢰성 있는 인식 결과를 기대할 수 없다.

또한 모폴로지를 이용한 방법[5], 인식기와 결합한 방법[6], 색상 정보를 드롭 아웃(Drop-out) 시키는 방법[7]들도 있다. 모폴로지 연산을 이용해서 선을 제거하는 방법은 구현은 간단하지만 문자 획이 끊어진 부분에서 복원이 부정확하게 되는 단점이 있다[5]. 인식기와 결합한 방법은 1차 복원 결과를 인식한 후 다시 복원 후보를 결정하는 과정으로 인해서 처리 시간이 증가하는 단점과 인식 결과의 신뢰성을 확신할 수 없는 문제점이 있다[6]. 인쇄된 문서 양식의 색상 정보를 드롭 아웃 시키는 방법은 원본 문서의 훼손이라는 이유로 현업에서 사용되고 있지 않는 실정이다[7].

문자열 복원은 선을 먼저 제거한 후 나머지 이미지에서 문자의 획을 복원하는 방법과[5-9] 선을 제거하기 전에 복원에 필요한 정보를 먼저 찾는 후에 선을 제거하며 복원하는 방법으로[2,10-14] 분류할 수 있다. 전자의 방법은 선을 제거할 때 문자 획의 일부도 함께 제거되기 때문에 나머지 영상만으로 정확한 복원이 어려운 단점이 있다. 후자의 방법은 선과 접촉된 문자 획의 모든 접촉 형태를 먼저 분석한 후 선이라고 판단되는 부분만을 제거하며 선과 문자 획이 만나는 영역을 접촉 형태에 따라 복원한다[2,12,14]. 이 방법은 복원에 필요한 정보를 먼저 추출하기 때문에 이미지의 소실을 감소시킬 수 있어서 전자의 방법보다 신뢰성 높은 인식 결과를 기대할 수 있지만, 문자의 획이 끊어져 있거나 선에 문자 획이 완전히 포함되어 있는 경우에는 복원이 어려운 단점이 있다. 본 연구에서 제안하는 방법은 후자의 방법과 같이 선을 제거하기 전에 선과 접촉된 문자 획 영역을 유형별로 분류하고 각 유형에 따른 복원 방법을 제안하고자 한다. 이 과정에서 문자의 획이 선에 완전히 포함된 경우의 복원도 고려한다.

문자 획의 복원을 낱자 단위 또는 문자의 열 단위로 구분하여 고려할 수 있다. 선에 접촉된 문자의 획을 낱자 단위로 복원하는 방법은[10,11,13] 우선 문자열을 낱자로 분리해야 하는 어려움이 따른다. 특히 필기된 문자열에서는 인접한 문자 사이에 다양한 형태의 접촉이 발생하며, 문자의 폭이 일정하지 않기 때문에 정확하게 낱자로 분리하는 일이 어렵다. 따라서 낱자 단위로 복원을 수행하는 방법은 서식의 문자 기입이 낱자 형태의 박스로 분리된 경우에 적용하는 것이 바람직하다. 본 논문에서는 문자열 특히 숫자열을 낱자로 분리하지 않은 상태에서 복원하는 방법을 제안하여 낱자 분리 과정의 어려움을 생략하고, 열 단위에서 찾을 수 있는 정보들을 활용하여 복원의 정확성을 높이고자 한다.

본 논문에서 제안하는 방법은 1) 숫자열 이미지를 복원 대상으로 하며, 2) 숫자열 단위로 복원하며, 3) 선과 접촉된 숫자 획의 영역을 유형별로 분류한 후 각 유형에 적합한 복원 방법을 제안하며, 4) 숫자 획의 일부가 선에 완전히 포함된 경우에도 복원을 시도한다. 제안한 방법의 실효성을 확인하기 위해서 은행의 입출금전표, 신용카드 매출전표와 같은 현장 데이타와 NIST DB에서 추출한 숫자열 이미지를 함께 사용하여 복원 방법을 평가하였다. 한글 및 영어 문자열의 복원은 향후 연구에서 고려하고자 한다.

본 논문의 2장에서는 제안하는 복원 방법을 구체적으로 설명하며, 3장에서는 실험 환경과 결과를 분석한다. 4장에서 본 논문의 결론을 맺는다.

2. 제안하는 방법

본 논문에서 제안하는 복원 방법은 서식 문서 구조 분석이 이루어진 후 추출된 숫자열 이미지를 대상으로 한다. 서식 문서의 구조 분석으로부터 얻어지는 정보인 각 영역의 입력 형태(인쇄/필기), 각 영역에 포함되어 있는 선의 형태, 개수, 선 사이의 간격 등이 숫자 열 이미지와 함께 이미 제공된다고 가정한다.

본 논문에서 제안하는 방법은 그림 2와 같이 전처리, 접촉 유형의 분류 및 복원의 세 단계로 처리된다. 전처리 과정에서는 입력된 이진 숫자열 이미지에서 서식의 선의 위치를 파악하고 선 및 숫자 획의 두께를 각각 검출한다. 접촉 유형 분류 과정에서는 선을 추적하며 수직 런(Run) 또는 수평 런 길이를 분석하여 접촉된 영역을 검출하고 각 영역의 유형을 분류한다. 접촉 영역을 유형별로 분류하는 것은 각 유형에 따라 적합한 복원 방법을 적용하기 위함이다. 본 연구에서는 숫자 획이 선과의 접촉이 아닌 다른 형태의 잡영에 의해서 훼손된 획은

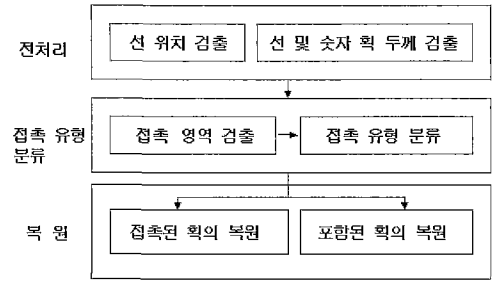


그림 2 제안하는 방법의 흐름도

고려하지 않는다. 복원 과정에서는 접촉된 유형에 따라 획이 선과 접촉된 경우와 획이 선에 포함된 경우로 나누어 복원을 수행한다. 선에 획이 접촉된 경우에는 비교적 쉽게 획을 복원할 수 있지만, 선에 획의 일부가 완전히 포함된 경우에는 주변의 접촉 유형과 선에 포함되지 않은 숫자 영역의 구조적인 특징을 분석하여 복원한다. 포함된 경우에는 필기된 숫자열의 복원과 인쇄된 숫자열의 복원으로 각각 나누어 제안한다.

2.1 전처리

2.1.1 선 위치 검출

선의 두께를 추정하고 문자 획과의 접촉 여부 및 영역을 판단하기 위해서 선의 위치를 찾는 과정이 필요하다. 선을 찾는데 허프 변환이 많이 사용되기도 하지만 계산 시간이 오래 걸리는 단점이 있다[3,11]. 본 논문에서는 선 위치를 찾기 위해서 히스토그램 프로파일을 이용한다. 이 방법은 선이 기울어져 있거나 두께가 얇은 경우, 또는 점선 등을 찾는데 어려움이 있지만, 적은 계산 시간으로도 비교적 쉽게 선을 찾는 장점이 있어서 서식 문서의 구조 분석에 많이 활용되고 있다. 본 논문에서는 선의 위치를 빠르고 정확하게 찾기 위해서 서식 문서의 등록 정보에서 쉽게 얻어지는 항목의 사전 정보인 영역에 포함된 선의 개수, 선의 형태, 선 사이의 간격 정보와 함께 히스토그램 프로파일을 분석한다. 그림 3과 같이 추출된 항목 이미지에서 수평선과 수직선의 개수가 일치할 때까지 히스토그램의 임계값을 조정하여 선의 위치를 추정한다. 이 방법을 이용하여 점선 또는 두께가 얇거나 끊어진 선도 쉽고 빠르게 찾을 수 있다.

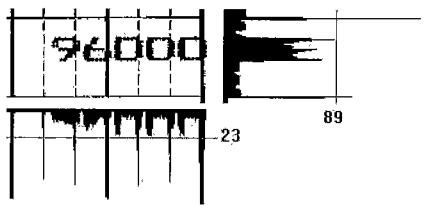
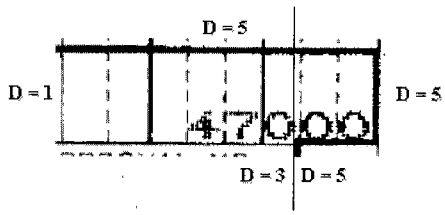


그림 3 항목의 사전 정보와 히스토그램 프로파일 분석을 이용한 선 위치 검출

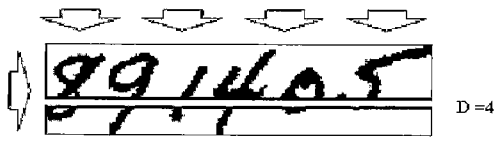
2.1.2 선 및 획 두께 검출

선과 숫자 획이 접촉된 영역을 찾고 복원을 정확하게 수행하기 위해서는 선 및 숫자 획의 두께를 검출하는 것이 필요하다. 검출 방법은 선을 추적하며 수직 또는 수평 런 길이를 측정하여 빈도수가 높은 런 값을 선의 두께로 결정한다. 수직선과 수평선 각각에 대하여 두께를 구하며, 두께가 변하는 선은 빈도수가 높은 두께 값들을 분류하여 각 선의 두께 값으로 결정한다. 그림 4(a)는 두께가 변하는 수평선에서 찾아진 두 개의 선 두께를 보여준다.

숫자의 획 두께는 가로 획과 세로 획에 대하여 각각 검출한다. 그림 4(b)와 같이 항목 영상에서 우선 선으로 판단된 부분을 가상으로 제거한 나머지 영상의 가로 및 세로 런 길이의 분포 가운데 빈도수가 높은 값을 각각 세로 획과 가로 획의 두께로 결정한다.



(a) 선 두께 검출



(b) 선을 가상으로 제거한 영역에서의 숫자 획 두께 검출

그림 4 선 및 획 두께 검출 예

2.2 접촉 유형 분류

2.2.1 접촉 영역 검출

본 본문에서는 숫자의 획이 수평선에 접촉되거나 포함된 경우를 대상으로 복원하는 방법을 제안하며, 수직선에 접촉되거나 포함된 경우는 같은 방법을 적용할 수 있기 때문에 설명을 생략한다. 우선 선과 획이 접촉된 영역을 찾기 위해서 선을 추적하며 수직 런 길이를 측정하며, 이 길이가 선의 두께보다 큰 영역을 접촉 가능한 영역으로 판단한다.

이와 같은 방식으로 접촉 영역을 찾으면 그림 5와 같이 한 두 픽셀의 두께 차이로 인해서 선과 획이 접촉된 영역으로 잘못 판단되는 영역도 포함된다. 따라서 접촉 영역으로 판단된 영역의 수평 런 길이가 짧고 수직 런 길이도 선의 두께보다 한 두 픽셀 정도만 큰 영역은 접촉 영역에서 제거시킨다.

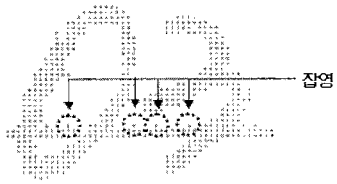


그림 5 잡영이 포함된 접촉 가능 영역

2.2.2 접촉 유형 분류

선과 획이 접촉된 영역을 선으로만 판단되어 제거할 영역, 선이 숫자 획의 일부로 판단되어 제거하지 않고 남겨야 하는 영역과 선을 제거하는 과정에서 발생할 수 있는 소실된 숫자 획을 복원해야 하는 영역으로 각각 분류한다. 따라서 선을 추적하며 각 세로 열을 아래의 세 가지 유형 중 하나로 분류하며, 이웃한 세로 열들이 동일한 유형으로 판단되면 결합하여 하나의 유형으로 결정한다. 이와 같이 분류된 영역은 그림 6과 같이 선의 위, 아래가 모두 접촉된 영역, 위 또는 아래 한 부분만 접촉된 영역과 위, 아래 두 부분이 모두 접촉되지 않은 영역으로 나뉜다.

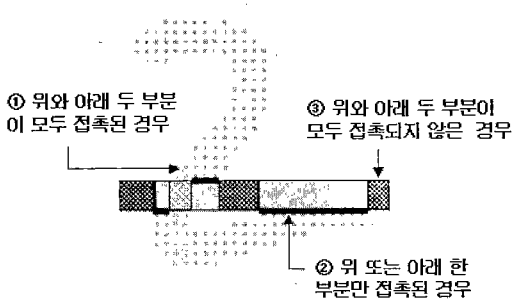


그림 6 접촉 영역의 분류

1) 교차 영역(CP: Crossing Part)

그림 6의 ①과 같이 선의 위, 아래 두 부분이 모두 획과 접촉되었다고 판단되는 영역으로서 교차 영역 또는 CP 영역으로 정한다. CP 영역은 선이 숫자 획의 일부이기 때문에 반드시 포함되어야 하는 영역으로서 제거하지 않는다. 따라서 복원할 필요가 없는 영역이다.

2) 복원할 영역(RP: Restorable Part)

그림 6의 ②와 같이 위 또는 아래의 한 부분만 접촉된 영역으로서 숫자의 획과 선의 일부분이 함께 존재하는 영역이다. 이 영역을 복원할 영역 또는 RP 영역으로 정한다. RP 영역의 복원은 숫자의 획에 속하는 부분은 남겨두며 선을 부분적으로 제거한다. 선의 제거와 획의 복원을 원활하게 수행하기 위해서 다시 이 영역을 각 세로 열의 수직 런 길이에 따라 다시 두 가지 유형으로 나눈다. 수직 런의 길이가 선의 두께보다 크고 선과 숫자의 획 두께를 더한 값보다 작거나 같은 영역을 RP1 영역으로 정하며, 수직 런의 길이가 선과 숫자의 획 두께를 더한 값보다 큰 영역을 RP2 영역으로 정한다. 그림 7은 위의 분류 기준에 따른 RP1 영역과 RP2 영역을 보여주며, 이 영역들에서 선의 일부인 회색 부분만을 정확히 제거하는 것이 복원 과정에서 할 일이다.

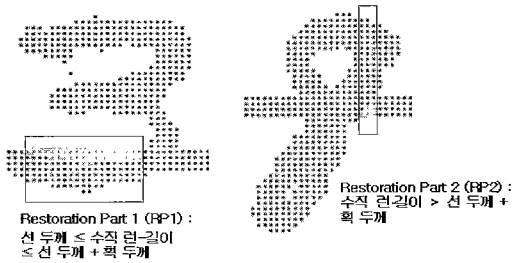


그림 7 RP1 및 RP2 유형

3) 모호한 영역(AP: Ambiguous Part)

그림 8과 같이 선의 위, 아래 두 부분이 모두 접촉되지 않은 영역으로서 이 영역은 단순히 선의 일부이거나 획이 선에 완전히 포함된 경우로 판단할 수 있기 때문에 모호한 영역 또는 AP 영역으로 정한다. 이 영역을 처리하기 위해서는 획이 선에 포함된 영역과 숫자와 숫자를 연결하는 선 영역을 구분하는 과정이 먼저 필요하다.

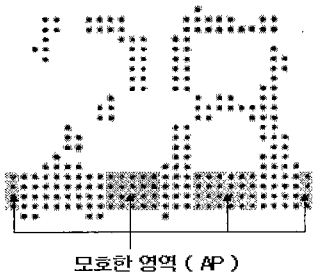


그림 8 모호한 영역

2.3 복원

복원은 숫자의 획과 선의 일부분이 함께 존재하는 영역을 찾아 숫자의 획이 원형대로 보존되도록 선의 일부

를 제거하는 과정이다. 획이 선이 교차하는 CP 영역은 모두 획에 속하는 영역이기 때문에 제거 및 복원을 고려할 필요가 없다. RP 영역은 RP1 영역과 RP2 영역으로 각각 나누어 복원한다. RP1 영역은 세로 열의 수직 런 길이에서 숫자 획의 두께를 남겨둔 나머지 선 부분을 제거하며 복원한다. RP2 영역은 선의 아래, 위 두 변과 숫자의 획이 만나는 두 점을 연결하는 직선을 구하여 그 직선을 기준으로 숫자인 부분과 숫자가 아닌 부분을 분리하여 복원한다. 수직 런 길이가 선의 두께보다 작거나 같은 모호한 영역인 AP 영역은 필기된 숫자열과 인쇄된 숫자열로 각각 구분하여 복원한다.

2.3.1 RP 영역의 복원

그림 9와 같이 RP1 영역은 숫자 획의 일부만이 선에 포함되어 평행하게 접촉된 영역으로서 획이 선의 위 또는 아래의 한 부분과 접촉된다. 이 영역의 수직 런 길이는 선의 두께보다는 크고 선과 숫자의 획 두께를 더한 값보다 작다. 복원은 수직 런 길이에서 숫자의 획 두께를 뺀 만큼을 숫자의 획과 접촉되지 않은 선 부분으로부터 픽셀 단위로 제거함으로서 이루어진다. 그림 9는 획 두께가 5일 때 RP1 영역의 복원 과정을 보여준다.

그림 9 RP1 영역의 복원

RP1 영역의 복원 방법을 그림 10(a)와 같이 숫자의 획 두께가 선의 두께보다 두꺼운 경우에 적용하면 그림 10(b)의 복원 결과가 만들어진다. 따라서 보다 정확한 복원을 위해서 RP1 영역 주변의 영역 정보를 함께 활용한다. 이러한 경우 대부분 그림 10(c)와 같이 영역이 AP, RP1, CP, RP1, AP로 분포되기 때문에 RP1의 모서리 흰색 점을 연결하는 직선을 구하여 숫자 획 부분과 선에 해당되는 부분으로 나누어 복원한다. 복원된 결과 이미지는 그림 10(d)와 같다.

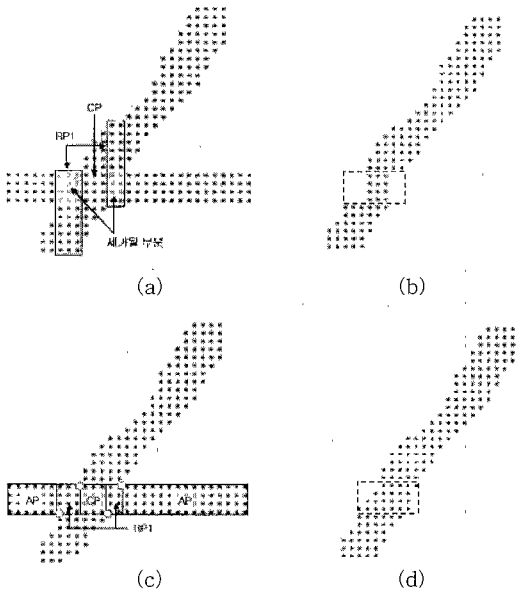


그림 10 획 두께가 선 두께 보다 큰 RP1 영역의 복원

RP2 영역도 숫자의 획이 선의 위 또는 아래 한 부분만 접촉되지만 수직 런 길이가 선과 숫자의 획 두께를 더한 값 보다 큰 영역이다. 이 영역의 복원은 선의 테두리 변과 획이 만나는 교차점을 연결하는 직선을 구하여 그 직선에 의해 숫자의 획 영역과 선 영역을 분리함으로서 이루어진다. 그림 11은 RP2 영역의 복원 결과이며, 그림 11(b)와 같은 경우에는 단지 RP2 영역을 제거한다. 이 경우 정확한 복원이 이루어진 것은 아니지만 복원 결과가 문자 인식 결과의 신뢰성을 낮추지는 않는다.

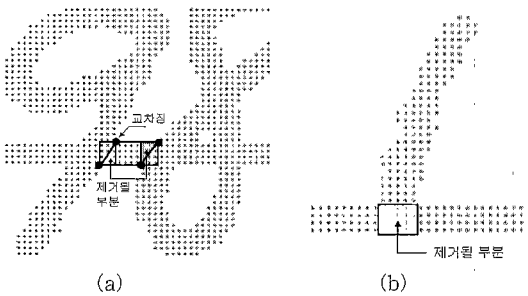


그림 11 RP2 영역의 복원

2.3.2 AP 영역의 복원

숫자의 획이 선에 완전히 포함되거나 접촉되지 않은

경우에는 수직 런 길이의 파악만으로는 선만의 영역인지 또는 획이 포함된 선의 영역인지를 파악할 수가 없다. 이와 같은 경우를 처리하고자 필기된 숫자열과 인쇄된 숫자열을 구분하여 각각에 적합한 복원 방법을 제안한다. 제안하는 방법은 경험에 의한 방법으로서 전반적인 결과의 향상에 초점을 두고 있음을 상기시키고자 한다.

사람의 필기 습관으로 인해서 하나의 필기된 숫자열의 위 또는 아래 부분에 있는 수평 획들이 동시에 같은 선에 접촉되는 경우는 거의 드물다는 특징을 이용한다. 제안하는 방법은 필기된 숫자열 AP 영역의 복원을 위해서 AP 영역 주변의 접촉 유형 분포를 분석하여 숫자 획의 포함 여부를 판단하여 복원한다.

서식에 인쇄되는 숫자열은 인쇄 특성상 숫자열의 위 또는 아래 부분에 있는 수평 획들 모두가 동일한 선에 포함될 수 있다. 제안하는 방법은 숫자열이 선에 완전히 포함되었는지를 먼저 판별하고, 포함되었다고 판단되면 숫자열을 낱자 단위로 분리한 후 개별 이미지의 구조적인 특징 정보를 활용하여 복원한다.

1) 필기된 숫자열의 복원

복원 과정은 다음과 같이 1) AP 영역의 폭을 파악, 2) AP 영역 좌우의 접촉 유형 파악, 3) 복원 결과 개선의 3단계로 진행된다.

첫 번째 단계에서 AP 영역의 폭을 파악하는 이유는 다음과 같다. AP 영역은 숫자의 획이 선에 완전히 포함된 경우도 있지만 숫자와 숫자 사이를 단순하게 연결하는 경우도 있다. 필기된 숫자열들을 관찰하면 필기되었기 때문에 선에 완전히 포함된 숫자 획의 길이가 길게 나타나는 경우는 드물다. 따라서 획이 선에 완전히 포함된 AP 영역과 숫자와 숫자 사이를 단순하게 연결하는 AP 영역을 구분하기 위해서 경험적으로 얻어진 AP 영역의 폭을 사용한다. 즉 AP 영역의 폭이 숫자 획의 두께보다 작거나 같은 경우에 대해서만 AP 영역 주변의 접촉 유형을 파악하여 복원한다. 반면에 AP 영역의 폭이 숫자 획의 두께보다 큰 경우에는 이 영역을 숫자와 숫자 사이를 연결하는 선으로 판단하여 제거한다.

앞의 과정에서 AP 영역의 폭이 정해진 값보다 작은 경우 다음 과정으로서 이 영역의 좌우에 나타나는 접촉 유형을 확인한다. 표 1은 숫자 획이 선에 포함된 경우에 AP 영역의 좌우에 나타나는 영역 분포를 경험적으로 얻은 결과이다. AP 영역의 폭이 숫자 획의 두께보다 작고 좌우 접촉 유형이 표 1에 포함된 경우에 대해서만 AP 영역을 숫자 획을 포함하는 영역 후보로 생각하여 복원을 진행한다. 표 1에 포함되어 있지 않은 좌우 주변의 영역에 대해서는 순수한 선의 일부로 판단하여 AP 영역을 제거한다.

표 1 주변의 접촉 유형에 따른 AP 영역의 복원

CP	AP	RP1
CP	AP	RP2
RP1	AP	CP
RP1	AP	RP1
RP1	AP	RP2
RP2	AP	CP
RP2	AP	RP1

끝으로 그림 12(a)와 같이 선의 두께가 숫자 획 두께보다 훨씬 큰 AP 영역에서는 영역의 두께를 조정하여 복원 결과를 개선한다. 두께 조정은 AP 영역 좌우가 모두 RP 영역인 경우에 수행한다. 그림 12(a)처럼 AP 영역의 좌우 영역이 RP1 영역이며 숫자의 획보다 선의 두께가 두꺼운 경우 좌우의 RP1 영역을 먼저 복원하여 얻은 흰색 점과 AP 영역의 모서리인 회색 점을 연결하는 사각형을 만들어서, 이 영역을 제외한 나머지 부분을 제거함으로서 두께를 조정한다. 그림 12(b)는 복원된 결과를 보여주며, 이 과정은 인쇄된 숫자열에서의 AP 영역의 복원에도 적용된다.

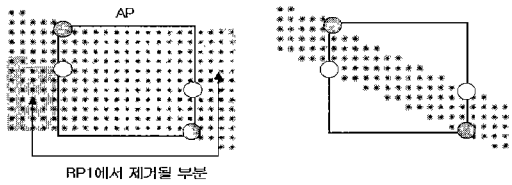


그림 12 AP 영역의 두께 조정에 의한 복원 결과의 개선

2) 인쇄된 숫자열의 복원

열 단위로 인쇄된 숫자열은 하나의 숫자가 서식 문서의 선에 포함되면 그 열의 나머지 숫자들도 대부분 선에 포함된다. 본 논문에서는 연구 결과의 실용적인 활용을 고려하여 200DPI의 저해상도에서 스캔된 낮은 수준의 신용카드 매출전표 이미지를 대상으로 인쇄된 숫자열의 복원 방법을 제안한다. 제안하는 방법을 요약하면 1) 우선 숫자열의 획들이 선에 포함되었는지를 판단하고, 2) 포함되었다고 판단된 숫자열을 낱자 단위로 분리

하며, 3) 각 낱자 이미지에서 선을 제외한 나머지 영상의 구조적인 정보를 파악하여 복원한다. 구조적인 특징 정보를 사용하는 것은 인쇄된 숫자들의 형태 정보가 거의 일정하기 때문이다.

첫 번째 과정에서는 획들이 선에 포함되었는지를 판별한다. AP 영역의 수직 런 길이가 측정된 선의 두께와 같고, 접촉 유형이 위 또는 아래의 한쪽 방향으로만 RP 영역과 AP 영역이 검출되며 CP 영역이 검출되지 않으면 획들이 선에 포함된 경우로 판단한다. 획들이 선에 포함되지 않고 접촉되기만 한 경우에는 RP 영역의 복원 방법을 적용하여 복원한다. 그림 13(a)는 인쇄된 숫자열의 획들이 선에 완전히 포함된 경우를 보여준다.

두 번째 과정에서는 선에 획들이 포함된 숫자열을 낱자 단위로 분리한다. 인쇄된 숫자의 문자 폭은 일정하기 때문에 이 특징과 선을 제외시킨 나머지 이미지의 히스토그램 프로파일을 이용하여 낱자 단위로 분리한다. 그림 13(b)는 선을 제외시킨 이미지의 수직 히스토그램 프로파일이며, 분리된 결과 이미지는 그림 13(c)와 같다. 그림 14는 위의 결과 이미지에서 획 두께도 조정된 결과로서 선의 수직 런 길이가 숫자의 획 두께보다 크기 때문에 선과 획의 두께 차이만큼을 선에서 깎아서 조정한 결과이다.

그림 13 선에 포함된 인쇄된 숫자열과 낱자 분리

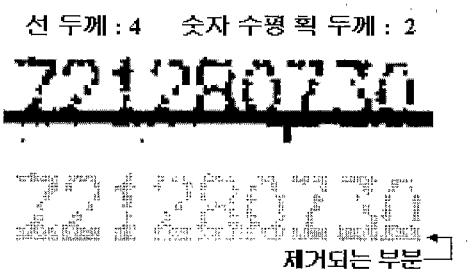


그림 14 숫자 획의 두께 조정

세 번째 과정에서는 선을 제외시킨 나머지 이미지에서 구조적인 정보를 파악하여 낱자 단위로 복원한다. 숫자열의 아래 부분의 획들이 선에 완전히 포함된 경우 구조적인 특징 정보를 파악하여 숫자를 0, 2, 3, 5, 6, 8 그룹과 1, 4, 7, 9 그룹으로 분류한다. 첫 번째 그룹의 숫자들은 남겨둠으로서 복원되며, 두 번째 그룹의 숫자들은 선과 숫자의 획이 접촉하는 형태를 다시 파악하여 복원한다. 숫자열의 위 부분이 선에 포함된 경우에는 숫자를 0, 2, 3, 5, 7, 8, 9 그룹과 1, 4, 6 그룹으로 나누어 아래와 비슷한 방법을 적용하여 복원한다.

이 과정을 구체적으로 설명하면 다음과 같다. 현재 복원하려는 숫자를 두 개의 그룹으로 분류하여 복원하기 위해서 각 영상에서 세 개의 구조적인 특징 정보를 파악한다. 첫 번째 특징은 선과 접촉된 획의 개수이다. 그림 15(a)와 같이 선과 접촉된 획의 개수가 하나이면 낱자 영상은 1, 2, 4, 7, 9에 속하며, 선과 접촉된 획의 개수가 둘이면 숫자 0, 3, 5, 6, 8에 속한다. 0, 3, 5, 6, 8은 모두 같은 그룹에 속하기 때문에 획을 남겨둠으로서 복원이 끝난다. 두 번째 특징은 그림 15(b)와 같이 낱자 영상의 가운데 영역에서 긴 수직 획의 존재 여부이다. 숫자 1, 2, 4, 7, 9중에서 숫자 1과 4만이 가운데 영역에 긴 수직 획을 갖기 때문에 긴 수직 획이 존재하면 숫자를 1 또는 4라고 판정하고 숫자의 획과 선이 접촉하는 위치까지 선을 제거하여 복원한다. 끝으로 세 번째 특징은 그림 15(c)와 같이 숫자의 상단에서 폐곡선(Hole)의 존재 여부이다.

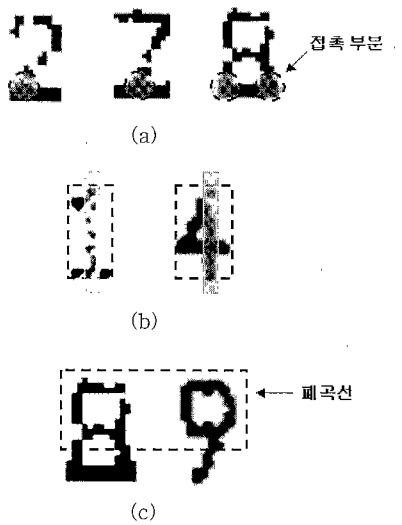


그림 15 선을 제외시킨 이미지에서의 구조적인 특징 추출

숫자 9는 폐곡선을 포함하기 때문에 숫자 2, 7과 쉽게 구분되며, 숫자 9로 판단되면 획과 선이 접촉하는 위치까지 선을 제거하여 복원한다. 그러나 최종적으로 남은 숫자 2와 7은 선을 전부 제거한 이미지와 전부 남겨둔 이미지를 복수의 후보 이미지로 제안한다.

그림 16은 위에서 설명한 각 과정의 결과 이미지를 보여준다. 숫자가 2 또는 7로 판정된 경우는 복수의 후보 이미지를 제시한 결과이다. 결과 이미지에서 숫자의 획이 선과 접촉되는 부분까지 정교하게 복원하지 않은 이유는 인식에 별 어려움이 없었기 때문이다.

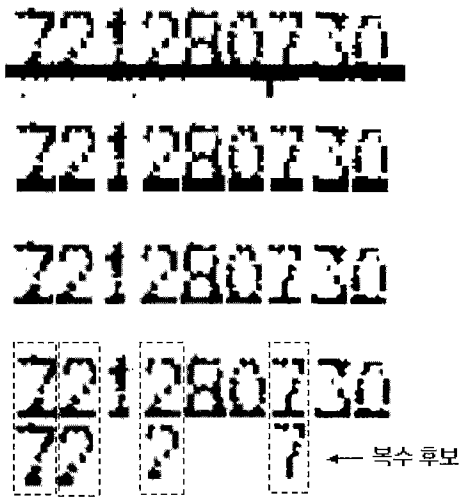


그림 16 선에 획이 포함된 인쇄 숫자 이미지의 복원

3. 실험 결과 및 분석

3.1 실험 환경

제안하는 방법을 Pentium II 300MHz PC에서 Visual C++ 6.0을 사용하여 구현하였다. 본 연구실에서 사전에 개발한 서식문서 구조분석 프로그램을 사용하여 영상의 기울어짐을 자동으로 교정하였으며, 추출된 각 항목 이미지에서의 선의 종류 및 개수 정보와 필기 또는 인쇄의 문자열 정보 등도 자동으로 추출된다.

실험은 현업에서 사용하는 200DPI의 저해상도에서 취득한 이미지를 주로 사용하였다. 필기된 숫자열의 복원 결과를 확인하기 위해서 은행 입출금전표에서 추출한 항목 이미지와 NIST DB #3에서 추출한 500개의 항목 이미지에 임의의 수평선을 추가하여 사용하였다. NIST DB #3의 각 필기 숫자열 이미지는 2개에서 10개의 숫자들로 구성되며 서로 인접한 숫자들이 접촉되어 있기도 하다. 인쇄된 숫자열의 복원 결과를 확인하기 위해서 신용카드 매출전표 이미지에서 추출된 저 수준의 항목 이미지를 사용하였다.

입출금전표와 신용카드 매출전표 이미지에 입력된 숫자열 이미지에 대해서는 사람의 시각적인 판단에 의하여 정성적으로 복원 결과를 평가하였으며, NIST DB의 숫자열 이미지에 대해서는 복원하기 전과 복원한 후의 이미지를 각각 인식하여 제안하는 방법의 타당성을 검토하였다. 숫자 인식은 Kirsh mask 특징과 Gradient 특징을 추출하여 학습한 두 개의 신경망 인식기를 결합하여 사용하였다[15].

3.2 결과 및 분석

그림 17은 은행 입출금 전표 이미지에서 추출한 숫자열의 복원 결과이다. 제안한 방법을 적용하여 필기된 이미지에 근접한 복원 결과를 얻을 수 있었다. 결과를 분석해보면 우선 CP 영역은 숫자열의 여러 부분에서 나타나고 있으며 이 영역은 그냥 남겨둠으로서 복원된다. 숫자열 '034254'에 포함된 숫자 '2'는 RP1 영역을 포함하고 있으며 제안한 방법을 적용하여 획의 끝이 비교적 정확하게 복원된 것을 확인할 수 있다. 또한 같은 숫자열의 숫자 '3'과 '5'는 RP2 영역을 포함하고 있기 때문에 직선의 방정식을 구하여 선과 숫자의 획을 구분함으로서 복원하였다. 그림 17에는 숫자의 획이 선에 완전히 포함된 AP 영역은 나타나지 않았으며 선으로 구성된 AP 영역만이 나타났다. 제안한 필기 숫자열 복원 방법에 의해서 이러한 AP 영역을 쉽게 제거하여 이웃하는 숫자에 끈(Ligature)으로 연결되지 않았다. 또한 그림 17에는 제안한 방법을 필기된 한글열 이미지에 적용한 결과도 보여주며 수직선과의 접촉에도 비교적 정확한 복원 결과를 확인할 수 있었다.

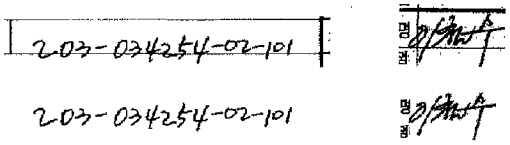


그림 17 필기된 숫자열과 한글 문자열의 복원 결과

그림 18은 NIST DB의 필기된 숫자 열에 인위적으로 수평선을 추가한 후에 제안한 방법을 적용하여 복원한 결과이다. 결과에서 볼 수 있듯이 NIST DB의 필기된 숫자열에 대해서도 만족스러운 복원 결과를 얻었다. 또한 복원 결과를 정량적으로 평가하기 위해서 선을 추가하기 전의 숫자 열 데이타들과 인위적으로 선을 추가한

후에 복원한 데이타들을 각각 인식하였다. 전체 500개의 숫자 열에 2,312개의 낱자 이미지가 있으며, 선이 추가되기 전의 원 이미지에 대해서는 85.2% 인식률을 얻었으며, 인위적으로 선을 추가한 후 다시 복원한 이미지에 대해서는 83.4%의 인식률을 얻어서 복원의 정확성을 어느 정도 확인할 수 있었다. 숫자 인식률이 전반적으로 낮은 이유는 숫자 열을 인식하는 과정에서 숫자와 숫자 사이의 접촉에 의한 분할 오류가 많았기 때문이다.

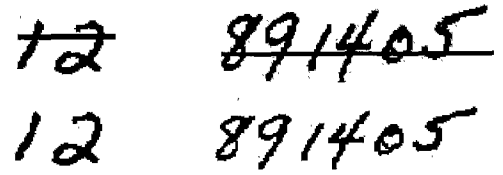


그림 18 NIST DB의 필기된 숫자열 복원

필기된 숫자열의 복원에서 AP 영역의 폭이 숫자 획의 두께보다 작은 경우에만 AP 영역을 복원하기 때문에 획의 두께보다 큰 AP 영역이 숫자의 수평 획을 포함하는 경우에는 정확한 복원 결과를 만들지 못한다. 그림 19는 NIST 데이타에 인위적으로 선을 삽입한 후에 복원한 결과로서, 숫자 '4'의 AP 영역의 폭이 7이며 숫자 획의 평균 두께가 5이기 때문에 AP 영역을 선의 일부로 판단하여 획을 삭제한 결과를 만들었다. 이러한 문제점을 해결하기 위해서 모든 AP 영역을 복수의 후보로 제시하여 인식에서 결정하도록 할 수 있지만 조합 가능한 숫자가 너무 많아지기 때문에 본 연구에서 고려하지 않았다.

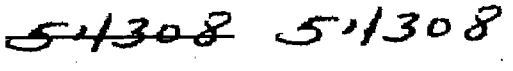


그림 19 선에 포함된 필기 숫자열 복원의 오류

인쇄된 숫자열 이미지의 복원 결과가 그림 20에 보여주고 있다. 그림 20(a)는 선과 숫자열이 접촉된 경우로서 RP1 영역의 복원 방법을 적용하여 원 이미지와 거의 비슷하게 복원되었다. 그림 20(b)는 각 숫자가 서식의 수직 실선 또는 점선과 접촉된 경우의 결과로서 수직선과 접촉된 경우에는 제안한 방법을 낱자 단위로 적용하여 복원하였다. 그림 20(c)는 숫자열의 아래 부분에 있는 획들이 서식의 선에 완전히 포함된 경우로서 숫자의 수평 획 두께 정보가 반영되어 복원된 결과이다. 숫자 '7'은 두 개의 후보를 복원 결과로 만들었다.

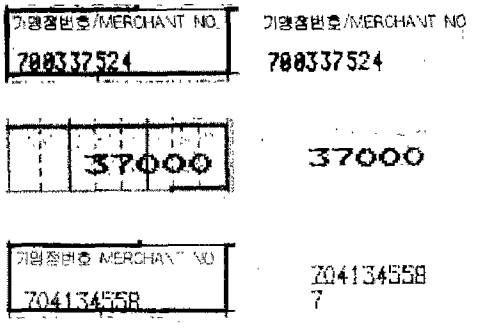


그림 20 인쇄된 숫자열의 복원

제안한 방법의 단점은 숫자 획의 두께를 구하는 과정에서 숫자열 이미지 전체를 대상으로 빈도수가 가장 많은 수평 및 수직 획의 두께를 각각 숫자 획의 두께로 검출하기 때문에, 낱자와 낱자 사이의 획의 두께 차이가 크면 정교하게 복원하지 못하는 것이다. 낱자마다 획 두께를 검출하여 정교한 복원을 할 수 있지만, 이 과정은 문자열을 먼저 낱자로 분리해야 하는 어려움이 따른다. 추후에 개선되어야 할 문제이다.

선과 접촉된 유형에 따라 복원 결과를 분석하면 다음과 같다. RP1, RP2 영역의 복원은 숫자의 획 두께가 일정하지 않은 경우 원 영상에 가까운 복원이 어렵지만, 이 결과가 인식에 크게 영향을 미치지 않는다는 것을 NIST DB를 이용한 실험에서 확인할 수 있었다. RP1 영역의 복원을 먼저 수행한 후에 RP2 영역의 복원을 수행하면, RP1 영역의 복원에서 얻은 획 두께 정보를 사용하여 좋은 복원 결과를 얻을 수 있었다. 인쇄된 숫자열의 AP 영역의 복원은 입력 영상의 질(quality)에 따라 결과가 크게 좌우되었다. Dot-Metrics 방식으로 인쇄된 숫자열에서 숫자의 구조적인 특징 정보를 정확하게 추출하는 것은 어려운 일이기 때문에 전처리 과정에서 이미지 향상 방법 등을 적용하여 질을 개선하는 과정이 필요하다.

4. 결론 및 향후 연구

본 논문에서는 서식 문서의 선에 접촉되거나 포함된 숫자열의 획들을 원래의 상태로 복원하는 방법을 제안하였다. 제안한 방법은 숫자열을 낱자 단위로 분리하지 않은 상태에서 먼저 접촉 유형을 분류한 후 각 유형에 적합한 복원 방법을 제안하였으며, 숫자의 획이 선에 완

전히 포함된 경우에도 복원하였다. 제안한 방법을 다양한 종류의 전표 이미지에서 추출된 숫자열과 NIST DB의 필기된 숫자열에 적용하여 현업에 사용할 수 있는 수준의 복원 결과를 확인할 수 있었다.

향후 연구로는 보다 정교한 복원 결과를 만들기 위해서 제안한 방법을 개선하는 것과 한글 및 영어 문자열 이미지에도 원활하게 적용할 수 있도록 개선하는 것이다.

참 고 문 헌

[1] S. Mori, C. Y. Suen, K. Yamamoto, "Historical Review of OCR Research and Development," Proc. of IEEE, Vol. 80, No. 7, pp. 1029-1058, July 1992.

[2] 유진용, 권영빈, "인쇄양식위에 기록한 필기문서의 라인제거 및 문자복원", 한국정보과학회 봄 학술발표논문집, 23권 1호, pp. 289-292, 1996.

[3] J. M. Gloger, "Use of Hough Transform to Separate Merged Text/Graphics in Forms," Proc. of 11th International Conference on Pattern Recognition, Vol. 2, pp. 268-271, 1992.

[4] O. Hori, D. S. Doermann, "Robust Table-form Structure Analysis Based on Box-reasoning," Proc. of 3rd International Conference on Document Analysis and Recognition, Vol. 2, pp. 218-221, 1995.

[5] D. Guillevic, C. Y. Suen, "Cursive Script Recognition: A Fast Reader Scheme," Proc. of 2nd International Conference on Document Analysis and Recognition, pp. 311-314, 1993.

[6] D. Wang, S. N. Srihari, "Analysis of Form Images," International Journal of Pattern Recognition and Artificial Intelligence, Vol. 8, No. 5, pp. 1031-1052, 1994.

[7] B. Yu, A. K. Jain, "A Form Dropout System," Proc. of 13th International Conference on Pattern Recognition, Vol. 3, pp. 701-705, 1996.

[8] S. Naoi, M. Yabuki, A. Asakawa, Y. Hotta, "Global Interpretation in the Segmentation of Handwritten Characters Overlapping a Border," IEICE Transactions of Information and Systems, Vol. E78-D, No. 7, pp. 909-916, 1995.

[9] S. N. Srihari, V. Govindaraju, A. Shekhawat, "Interpretation of Handwritten Addresses in US Mailstream," Proc. of 2nd International Conference on Document Analysis and Recognition, pp. 291-294, 1993.

[10] Y. T. Chung, K. Y. Lee, J. H. Paik, Y. B. Lee, "Extraction and Restoration of Digits Touching or Overlapping Lines," Proc. of 13th International Conference on Pattern Recognition, Vol. 3, pp. 155-159, 1996.

[11] Y. H. Tseng, H. J. Lee, "Interfered-character Recognition by Removing Interfering-lines and Adjusting Feature Weights", Proc. of 14th International Conference on Pattern Recognition, Vol. 2, pp. 1865-1867, 1998.

[12] J. Y. Yoo et. al. "Line Removal and Restoration of Handwritten Characters on the Form Documents," Proc. of 4th International Conference on Document Analysis and Recognition, pp. 128-131, Germany, 1997.

[13] 정영태, 이관용, 백종현, 이일병, 변혜란, "선에 걸친 숫자 영상의 추출 및 복원", 한국정보과학회 봄 학술발표논문집, 23권 1호, pp. 273-276, 1996.

[14] D. S. Doermann, A. Rosenfeld, "The Processing of Form Document," Proc. of 2nd International Conference on Document Analysis and Recognition, pp. 497-501, 1993.

[15] 백종현, 조성배, 이관용, 이일병, "이중 결합구조를 갖는 다중 인식기 시스템", 한국정보과학회 봄 학술발표논문집, 23권 1호, pp. 281-284, 1996.

이 창 현

1998년 순천향대학교 전산학과 학사. 2000년 연세대학교 컴퓨터과학과 석사. 2000년 2월 ~ 현재 삼성SDS 재직. 관심분야는 인공지능, 영상처리, 데이타마이닝 등

최 영 우

1985년 연세대학교 전자공학과 학사. 1986년 University of Southern California 컴퓨터공학과 석사. 1994년 University of Southern California 컴퓨터공학과 박사. 1994년 10월 ~ 1997년 2월 LG전자기술원 선임연구원. 1997년 3월 ~ 현재 숙명여자대학교 정보과학부 조교수, 부교수. 관심분야는 영상처리, 패턴인식, 문자인식 등

김 경 환

1984년 서강대학교 전자공학과 학사. 1986년 서강대학교 대학원 전자공학과 석사. 1996년 State University of New York at Buffalo 전기 및 컴퓨터 공학과 박사. 1986년 ~ 1991년 금성전기(정밀) 기술연구소 선임연구원. 1993년 ~ 1997년 CEDAR/SUNY at Buffalo Research Scientist. 1997년 ~ 현재 서강대학교 전자공학과 조교수. 관심분야는 문자 및 문서영상 처리, 영상신호해석, 패턴인식, 신경회로망, Embedded System Design 등

이 일 병

1976년 연세대학교 전자공학과 공학사. 1983년 University of Illinois 컴퓨터과학과 석사. 1986년 University of Massachusetts 컴퓨터과학과 박사. 1985년 ~ 현재 연세대학교 컴퓨터과학과 교수. 연세대학교 소프트웨어응용연구소 소장 역임, 한국정보과학회 인공지능연구회 회장 역임, 한국정보과학회 신경회로망연구회 회장 역임, 한국정보과학회 이사 역임, 한국인지과학회 이사 역임, 한국퍼지및지능시스템학회 이사 역임. 2001년 현재 한국인지과학회 회장, 한국데이타마이닝학회 부회장. 관심분야는 인공지능, 패턴인식, 생체인식, 인지과학, 데이타마이닝 등