論 文

명함에서 지식베이스를 이용한 구성요소의 추출

正會員 李 成 範* 正會員 南宮 在 贊**

The Component Extraction Using Knowledge-Base from Name-Card

Sung Bum Lee*, Jae Chan NamKung** *Regular Members*

要 約

본 논문은 명함에서 지식베이스를 이용하여, 정보항목를 자동적으로 추출하는 실험을 하였다. 본 연구에서 사용한 기본개념은 명함내에 지식으로 항목과 요소들간의 관련정보 및 구조적인 정보를 이용한다. 계층적인 지식을 지식베이스로 기술하기 위해 프레임표현을 사용하고, 명함에서 항목과 그룹후보를 추출하기위한 영역분류 알고리즘을 제안했다. 100개의 대상 명함에 대해서 실험한 결과는 95%이상의 추출율을 얻었다.

ABSTRACT

This paper presents the automatically extracting method of data item from name-cards using knowledge-base. In our approach, we utilize a structural information and a relational information between data items and elements with knowledge in the name-cards. To describe a hierarchical knowledge, we uses a frame structure and we propose an algorithim of domain classification to extract item and group candidate domains from the name-cards. From the experimental results, we obtain the extraction rate, 95%, for 100 samples.

I. 서 론

근래 문서인식에 관한 연구는 방대한 데이타량과 빠른 속도를 요구하는 관계로 연구 및 개발에 많은 어려움이 있어 왔으나, 최근에는 하드웨어 및 소프트웨어의 기술발달이 괄목하여 실용화 추세에 있다. 문서인식에 대한 최근 동향을 분석해보면, Wang D.와 Srihari S. N.은 신문을 중심으로 RLSA(Run Length Smoothing Algorithm)를 적용하여 문서의 영역화를 하였다.[1]

이 외에도 Tang Y. Y.등은 엔트로피 함수를 이용하여 하향식(top down), 상향식(bottom up) 접근과 문서구조를 이론적으로 분석하여 해석을 하였고, 문서서식지식과 문서서식 기술언어(DFDL)를 제안했다.[2]

S. M. Kerpedjiev는 원래 문서에서 문서분류처리를 하기위해서 프로토타입(prototype)시스템을 사용

*大宥工業專門大學
**光云大學校 電子計算機工學科
論文番號 : 93-123

하여 문서로부터 정보를 자동추출 하였다.[3] 또 J. Kreich등은 지식베이스 영상해석을 이용하여 스캔된 문서 래스터화상에서 문서해석과 번역을 위한 SODA (업무문서해석을 위한 시스템)의 실험장치로서 상업 통신문의 요소를 추출하였고[4], Chenevoy Y.등은 문서의 구조적 문맥을 고려한 가설방법론과 구조의 가설을 사용한 Graphein이라는 흑판기본 시스템을 써서 구조화된 문서를 인식하는 등 논문이 발표되었다.[5] 특히 영문자는 단순하다는 장점하에 이미 시스템으로 완성되어 사용자들을 만족 시켜주고 있고, 현재 이러한 시스템들은 MS-WINDOW 버젼으로 발표되고 있으며 다양한 파일 포멧을 임포트(import)하고 있다. 국내에서는 김진형등의 논문[6] 및 시스템은 SUN SPARC 2시스템에서 문서와 문자인식 시스템을 개발 시판하나, 고가이어서 일반 사용자가 이용하기에는 매우 불리한 여건에 있다.

본 연구는 PC 상에서 실행되는 문서인식 시스템을 개발하기위해 적은 데이타와 제약을 가진 문서의 한 형태인 명함을 택하여 한글과 국내명함에 알맞게 알고리즘을 개발하는 연구로서, 지금까지 문서형태로 저장되던 정보를 대량축적하고, 신속한 처리의 형태인 데이타-베이스(data-base)화를 실현하기위한 연구의 한 단계이다.

문서 데이타를 자동으로 데이타-베이스에 입력하기 위해서는 문서의 기술내용이 다양화되어 있는 것을 화상으로서가 아닌 문자코드로 축적되어야 한다. 이를 실현하기 위해서는 입력방법이 문제가 되는데 이를 위하여 입력항목을 분류하여 추출할 필요가 있게 된다.[7] 이때 문서의 구성요소는 그역할에 따라서 배치구조(layout)를 가지므로 이 배치구조를 근거로 하여 구성요소를 추출할 수 있다고 생각된다.[8,9]

그렇지만 배치구조를 이용하여 문서구조를 해석하는 경우, 배치구조가 대상문서에 의존하기 때문에 다양한 형태의 문서에서는 방법상에 문제가 있다. Yeh P.S.등의 방법에서는 배치구조를 처리 매개변수(parameter)로하여 순서적으로 표현했었다.[10] 그러나 이와같은 방법에서는 대상 화상의 고유 배치구조가 처리 절차내에 들어있기 때문에 대상 문서를 변경할 때 많은 노력이 필요하며 시스템의 범용성에 문제가 있다.

또한 Niyogi D.등의 규칙베이스(rule base system)에서는 규칙을 이용하여 배치구조를 표현하기 때문에 배치구조를 이룬 계층성이 명시적으로 표현되지 않는 등의 표현 능력에 문제가 있었다.[11]

본 논고에서는 이상의 논의를 근거로하여 지식베이스(knowledge base)를 이용하여 명함화상의 구성요소 추출을 행한다. 여기에서는 대상문서의 배치구조에 관한 지식, 논리구조에 관한 지식 및 배치구조와 논리구조의 대응관계에 관한 지식등을 지식베이스화하고 배치구조가 본질적으로 가지고 있는 논리구조를 프레임형식으로 전개하며, 지식베이스를 구축하는 기술을 프레임 술어을 정의하여 사람이 이해하기 쉬운 술어로 표현하였다. 또 기술을 할때 상대적, 계층적인 특징량 및 배치구조 술어를 사용하므로 기술의 용이성과 구조변화에 확장성을 기대하고 있다.

또 추출하기 위해서 각 문자, 항목 및 그룹의 블록화 해야하는 데, 종래 가록현등은 CRLCA를 이용하여 문서분할을 하였으나, 그 처리시간과 문자열 추출은 새로운 feature 추출과정을 거쳐야하는 등의 문제가 있었다.[12] 본 연구에서는 문자블록, 항목블록, 그룹블록의 순으로 각 영역정보를 블록화하는 알고리즘을 제안하여, 각 영역에 대한 지식 베이스의 정보를 얻도록하였다.

본고에서 실험 대상으로 선택한 명함은 가로쓰기와 세로쓰기형태가 있고 직업별로 그 배치구조가 다양하지만, 제약을 전산관련 및 학교에 근무하는 사람의 명함을 대상한다. 그 유효성을 검토하기 위해 이들자료의 특징을 조사해보면 80%이상이 가로쓰기 형태로 되어있어 이를 중심으로 논의를 전개토록하였다.

Ⅱ. 명함의 유형분석

명함을 직업에 관계없이 500장을 분류 정리하여 본 결과 405장(약 80%)이 가로쓰기이고 95장(약 20%)이 세로쓰기 형태였으며, 그 크기는 가로 및 세로쓰기형태에 관계없이 5.5＊9㎝로 일정하였고 직업에 따라서 그 배치가 달랐다. 예를 들면 영업직, 인테리어등의 업종은 그 배치가 아주 독특하고 변화가 많으나, 본고에서 대상으로한 교직, 기술직 및 관공서등은 거의 규격화되어, 아래와 같은 배치 특성이 있었다.

2-1. 가로쓰기 유형

가로쓰기 명함의 특징은 그룹으로 크게 3가지 즉 회사명, 이름, 주소그룹으로 되어있고 각 그룹은 몇개의 항목으로 이루워졌고 항목내에서는 문자크기가

거의 동일하며 그룹은 위에서 아래로, 항목은 위에서 아래와 좌우로 배치되어있다. 명함내의 구성요소들은 회사마크, 회사명, 소속부서명, 직함, 이름, 주소, 우편번호, 전화번호, 팩스번호, 텔랙스번호등이며, 내용을 기호, 문자, 숫자로 표현하였다.

기호로는 회사마크가 있고 문자로는 한글, 한자, 영자(뒷면기재 포함)이고, 숫자로는 아라비아 숫자이었다.

문자나 숫자의 크기는 대부분 회사명과 이름이 가장 크고 직함, 소속부서명, 주소등은 작은 크기의 문자나 숫자로 쓰여졌다.

이 유형의 배치구조는 상단에는 회사명그룹이 배치되어 있어, 좌측에 회사마크가 있고 우측에 회사명과 소속부서가 있으며, 중앙에는 이름그룹이 있어 좌측은 직함이 우측에는 이름이 있고, 하단에는 대개 조직의 연락처에 관한 정보로서 주소, 전화번호, 팩스번호 및 텔랙스번호등이 배치되어 있다.

2-2. 세로쓰기 유형

이 유형의 모든항목의 문자는 위에서 아래로 쓰여지고, 이들 항목의 모임인 회사명, 이름, 주소그룹은 우측에서 좌측으로 배치되어 있다.

논리구조는 가로쓰기 형태와 유사하나 배치구조만이 대칭적으로 다를 뿐이었다.

Ⅲ. 명함 영역지식(domain knowledge)

명함화상의 구조는 배치구조와 논리구조로 나눌 수 있다. 배치구조란 문자, 문자열, 문자열의 집합인 항목들이 공간에서 구성요소간에 배치와 포함등의 관계를 나타내는 것으로 시각적으로 이해될 수 있다. 예로서「문자열은 문자로 구성되고 좌에서 우로 쓰여졌다」라는 것은 구성요소간에 포함 및 배치관계를 나타낸다. 또 논리구조란 의미구조라고도 하며 예로서 목차와 본문등의 논리적인 구성요소의 포함, 등가, 관련등의 관계로 나타 낼수도 있다. 예로서「장(chapter)과 그에 따른 절」이라는 관계는 논리적 구성요소간의 포함관계를 나타낸다. 명함 화상인식의 목적은 명함화상에서 논리구조를 추출하는 것이다. 그런데 명함에는 명함내용의 관계를 어느 정도 이해할 수 있는 배치구조가 있으므로 배치구조와 논리구조와의 사이에 대응관계가 존재하게 된다.

따라서 이 대응관계를 이용하여 배치구조에 관한 지식으로부터 논리구조를 추출하는 것이 가능할 것이다.[13] 이를 실현하기 위해서 명함의 배치구조에 관한 지식, 명함의 논리구조에 관한 지식, 명함의 배치구조와 논리구조와 논리구조의 대응관계에 관한 지식이 필요하게 된다.

본 논문에서는 이들 3가지의 지식을 명함 화상인식에 있어서의 영역지식으로 하여, 논리적인 구성요소를 프레임이라 불리는 형태로 표현을 시도한다.

3-1. 명함화상에서의 배치구조

일반적인 명함화상에는 소유자에 관한 정보와 소유자가 소속한 조직에 관한 정보 및 조직의 연락처에 관한 정보를 갖고 있다.

이들을 각각 회사명그룹, 이름그룹, 주소그룹으로 크게 3가지로 분류하며, 또 각 그룹은 하위레벨의 구성요소로 항목을 갖는데 이들은 각각 회사명 그룹에는 마크, 회사명, 소속부서명 등의 항목이 있고, 이름그룹에는 직함과 이름의 항목이 있으며, 주소그룹에는 우편번호, 주소, 전화번호, 텔랙스번호등의 항목들이 각각 선택적으로 있으며 이항목들은 문자열의 일부 혹은 전부에 대응된다. 이들을 공간에 위에서 아래로 혹은 좌에서 우로 위치시킨 것은 배치구조라 부르고 각 그룹과 항목은 그림 1과 같이 각 구성요소가 배치되어 있다.

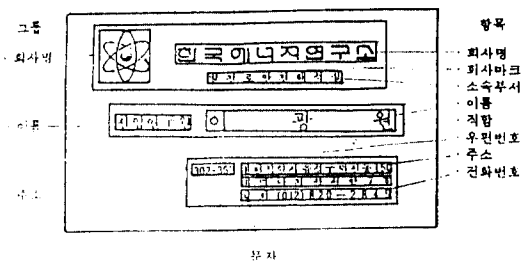


그림 1. 명함의 배치구조
Fig. 1. Layout structure of name-card.

3-2. 명함화상의 구성요소의 기술

배치구조를 지식으로 표현하는 경우에는 배치상에 의미없는 변동은 흡수하고 본질적인 구조를 표현하는 것이 필요하므로 구성요소의 배치단위인 문자열과 블록들의 영역은 구형을 기본으로 갖는 것으로 하며, 각 계층에서 구성요소가 명함화상 내에서 차지하는 영역을 구형으로 받아 들이기로 한다. 구형에 관한 기본 특징량으로는 그림 2에서와 같이 원점을 명

함화상의 좌상부로하고 가로축을 X, 세로축을 Y로 잡은 좌표계를 쓰면, 가로폭 Wx, 세로폭 Wy, 면적 A, 중심 C와 정방형도(square degree) S를 다음과 같이 정의한다.

$$Wx = Xe - Xb + 1$$
$$Wy = Ye - Yb + 1$$
$$A = Wx * Wy$$
$$C = (Cx, Cy)$$
$$Cx = (Xb + Xe)/2$$
$$Cy = (Yb + Ye)/2$$
$$S = 1 - Wy/Wx \quad : \quad Wy \Leftarrow Wx$$
$$= Wx/Wy - 1 \quad : \quad Wx \Leftarrow Wy$$

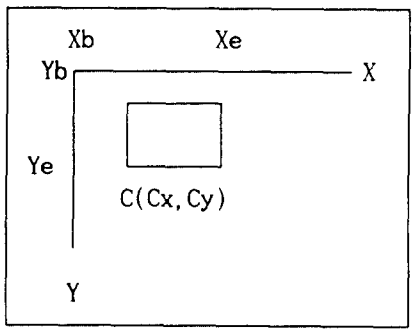


그림 2. 기본구형의 표현
Fig. 2. Representation of basic rectangle.

위 정의에서 특히 정방형도S는 구성요소인 구형이 정방형(정사각형)의 정도를 나타내는 지표로서 구형이 정방형에 가까우면 0에 가까우며, 구형이 세로로 긴 모양일 때는 −값을 갖고 구형이 가로로 긴모양을 가질 때는 +값을 가지게 된다.

그런데 위 정의의 특징량들 중에 정방형도를 제외한 가로폭, 세로폭, 면적과 중심은 화상의 크기에 의존하기 때문에 그대로 지식으로 기술하기에는 적합하지 않다. 따라서 본고에서는 명함화상 중에 존재하는 각 구성요소에 대하여 프레임으로 작성되도록 그 중에 정규화 특징량과 비교특징량에 대해서 술어로 바꾸워서 각 slot에 기술하기로 한다.

정규화 특징량으로는 부분-전체관계를 가진 구성요소들에 대하여 부분 구성요소의 기본특징량을 전체 구성요소의 기본특징량으로 정규화한 것을 다음과 같이 정의한다.

* 정규화 가로폭(normalized-horizontal-width) :
  $Wxn = Wxp / Wxu$
* 정규화 세로폭(normalized-vertical-width) :
  $Xyn = Wyp / Wyu$
* 정규화 면적(normalized-area) :
  $An = Ap / Au$
* 정규화 중심(normalized-center) :
  $Cn = (Cxn, Cyn)$
  Cxn(normalized-horizontal-center) $=$
  $2 * (Cxu - Cxp) / Wxu$
  Cyn(normalized-vertical-center) $=$
  $2 * (Cyu - Cyp) / Wyu$
* 정규화 좌표(normalized-coordinate)
  $D_{xb} = |X_b - X_{bu}| / W_{xu}$
  $D_{xe} = |X_e - X_{eu}| / W_{xu}$
  $D_{yb} = |Y_b - Y_{bu}| / W_{yu}$
  $D_{ye} = |Y_e - Y_{eu}| / W_{yu}$

여기서 첨자 u, p는 각각 전체, 부분의 구형 기본특징량을 나타낸다.

비교특징량으로는 비교구형 a, b사이에 비교특징량으로 다음과 같이 정의한다.

* 비교가로폭(compared-horizontal-width)

$$Wxc = \frac{Wxa - Wxb}{\max(Wxa\ ,\ Wxb)}$$

* 비교세로폭(compared-vertical-width)

$$Wyc = \frac{Wya - Wyb}{\max(Wya\ ,\ Wyb)}$$

* 비교면적(compared-area)

$$Ac = \frac{Aa - Ab}{\max(Aa\ ,\ Ab)}$$

* 비교중심(compared-center)
  * compared-horizontal-center ;
    $Cxc = 2 * (Cxa - Cxb) / Wxu$
  * compared-vertical-center ;
    $Cyc = 2 * (Cya - Cyb) / Wyu$
* 비교 좌표(compared-coordinate)
  $E_{xb} = 2 * (X_{ba} - X_{bb}) / (W_{xa} + W_{xb})$
  $E_{xe} = 2 * (X_{ea} - X_{eb}) / (W_{xa} + W_{xb})$
  $E_{yb} = 2 * (Y_{ba} - Y_{bb}) / (W_{ya} + W_{yb})$
  $E_{ye} = 2 * (Y_{ea} - Y_{eb}) / (W_{ya} + W_{yb})$

여기서 첨자 a, b는 각각 비교대상인 구형 A, B의 기본특징량을 구별하기 위한 것이다. 이상과 같이 정의한 구형의 특징량으로 프레임에 기술하기 위해 프레임술어를 만든다. 프레임술어는 약어와 지정자로 구성하도록 정의한다. 약어는 정규화 특징량, 비교특징량, 정방형도, 구성요소의 수를 기술하는 구성요소수(number-of)와 상대위치(position)들이 있다.

지정자로는 방향지정자, 정도지정자와 수량지정자를 고려하여, 상대위치를 기술하는 경우에 방향지정자로서 위(up), 아래(down), 좌(left), 우(right)의 4방향을 고려하고, 구성요소수를 기술하는 경우에는 수량지정자를 각각 사용하며, 정도지정자는 상대위치로서 프레임술어에 사용되며 각 프레임술어 마다 특징량의 구간을 치역한정자의 집합으로 지정할 수가 있다. 각 지역한정자는 각각 일정한 치역을 표현하고 있는데, 정도지정자의 정의는 J.H. Connell and M. Brady[14]의 Gray-code의 개념을 이용했다. Gray-code는 연속한 구간을 표현할 때에 구간폭을 임의로 선택하여 기술하며 또 구간의 포함관계를 코드로 용이하게 나타내는 이점을 갖고 있다.

이상과 같이 그림 3은 비교특징량에 대한 치역한정자의 정의 예를 표시하며, 정도지정자는 치역한정자의 집합으로 표시되고 정도지정자가 지정한 구간은 치역한정자의 집합의 공통구간이 된다. 정의의 정도지정자에 의해서 비교특징량의 예로 표 1와 같이 원하는 구간폭을 기술한다.

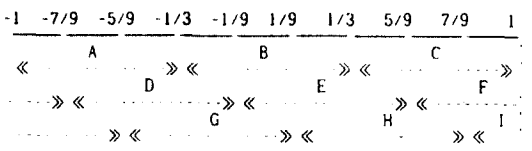


그림 3. 치역한 정자
Fig. 3. Domain definition.

표 1. 정도지정자
Table 1. Extent appointor

구간 폭	구 간	정도지정자
2 / 3	$-1/3 < P < 1/3$	{B}
4 / 9	$-1 < P < -5/9$	{A, I}
2 / 9	$-1/9 < P < 1/9$	{B, E, G}

이상을 조합하여 프레임에 기술하기 위해 프레임술어를 만든다. 프레임술어의 정의는 약어의 연언 혹은 약어와 정도지정자로서 표 2와 같다.

이상처럼 정규화 특징량, 비교특징량, 상대위치 및 구성요소수를 이용하여 구성요소의 영역을 기술하면 구형의 좌표로 직접기술하는 것에 비하여 간결한 기술이 가능하며 또 명함의 크기에 의존되지않을 수 있다는 장점이 있다.

프레임술어 ; [(약어)(지정자)]
지정자 ; [(방향지정자)(수량지정자)(정도지정자)]
방향지정자 ; [(up)(down)(left)(right)]
수량지정자 ; [(비교연산자) x, x는 자연수]
비교연산자 ; [(==)(≤)(〉=)]

표 2. 프레임술어의 정의
Table 2. Define of frame predicate.

프 레 임 술 어	정 의
horizontal-centering	{normalized-horizontal-center [B, E, G]}
vertical-centering	{normalized-vertical-center [B, E, G]}
upper-end	{ D_{yb} [A, F, I]}
bottom-end	{ D_{ye} [C, F, I]}
left-edge	{ D_{xb} [A, F, I]}
right-edge	{ D_{xe} [C, F, I]}
horizontal-alignment	{ E_{yb} [B]} and {E_{ye} [B]}
vertical-alignment	{ E_{xb} [B]} and {E_{xe} [B]}
vertical-right-indention	{ E_{yb} [0, ∞]} and not(horizonta-alignment)
vertical-left-indention	{ E_{yb} [−∞, 0]} and not(horizonta-alignment)

Ⅳ. 명함화상의 지식 표현 기술

4-1. 지식의 표현으로 프레임 기술형식

그림 1의 명함화상에서와 같이 각각의 문자가 차지한 영역을 구형으로 표현하고 또 일정 방향으로 나란히 놓인 문자열도 구형영역으로 표현되었다. 따라서 같은 성질과 방향을 가진 구성요소를 포함한 구형으로 고려하면, 그림 4와 같이 6 level의 논리구조를 갖는 것으로 계층구조로 도해할 수 있다. 명함구조의 계층성을 지식표현으로 할 수 있는 선언적 표현방법 중에서 그림 5와 같이 Minsky의 프레임을 사용하여 지식표현 능력을 향상 시켰다.

프레임에는 다음의 슬롯을 만들어서 그림 5의 프레임을 표현한다.

(1)NAME : 프레임 명 혹은 부 프레임 명을 기술한다.

(2)SELF : 구형요소의 특징들(구형의 숫자, 정방형도, 크기등)을 기술한다.

(3)PART-OF : 프레임에서 전체와 부분관계를 이룰때 하위의 프레임명을 기술한다.

(4)SIMILARITY : 비교 특징량인 프레임 술어를 기술한다.(표 2)

(5)RELATION : 정규화 특징량인 프레임 술어를 기술한다.(표 2)

(6)IS-A : 상위-하위관계를 이룰때 상위 프레임 명을 기술한다.

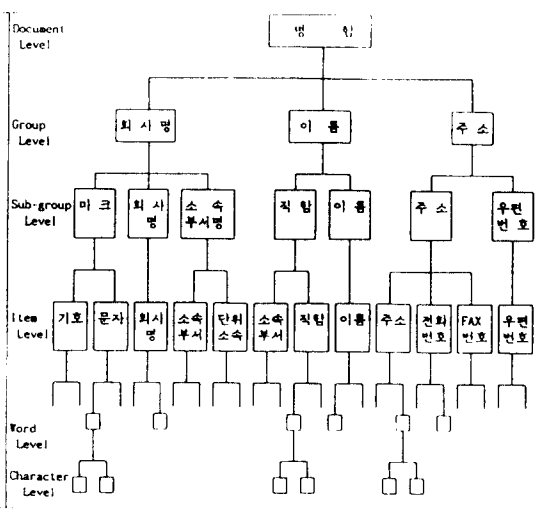


그림 4. 명함화상의 논리구조

Fig. 4. Logical structure of name-card.

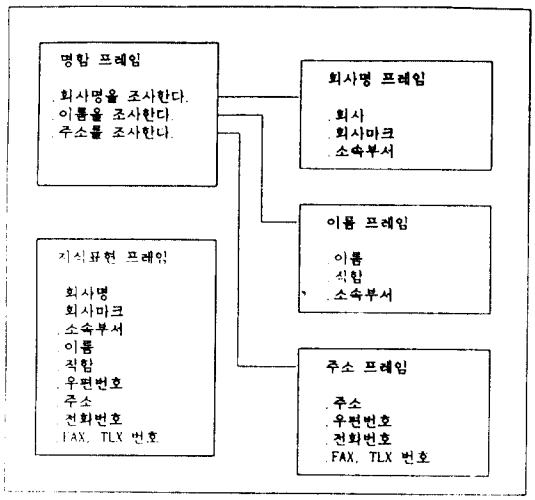


그림 5. 명함의 프레임전개

Fig. 5. Frame diagram of name-card.

4-2. 기술 예

그림 1에 명함화상중에 회사명 그룹에 대하여 지식베이스로 하기위해서 프레임으로 기술한 예를 표 3와 같이 기술할 수 있다.

표 3. 회사명그룹의 기술예

Table 3. Example of description for office-name-group.

FRAME : NAME	(Office-name-group)	
: SELF	(Number of = 1, Square-degree = +)	…… 1)
: PART-OF	(Mark-item, Office-name-item, Sub-office-name-item)	… 2)
: SIMILARITY	(Horizontal-alignment)	…… 3)
SUB-FRAME		
: NAME	(Mark-item)	…… 4)
: IS-A	(Office-name-group)	
: RELATION	(Left-edge, Upper-end)	… 5)
SUB-FRAME		
: NAME	(Office-name-item)	
: IS-A	(Office-name-group)	
: RELATION	(Upper-end)	
SUB-FRAME		
: NAME	(Sub-office-name-item)	
: IS-A	(Office-name-group)	
: RELATION	(Bottom-end)	

위 기술 예의 상세히 살펴보면

1) 구성요소의 특징을 나타내는 부분으로, 명함화상

내에 회사명그룹이 한개있고 그방형도가 +값으로 가로로 긴 직사각형인 모양을 가지고 있다.
2)레벨이 다른 구성요소사이에 부분-전체관계를 표현한 것으로 Mark-item, Office-name-item, Sub-office-name-item으로 구성되어 있는 것을 뜻한다.
3)구성요소의 비교관계를 나타내는 것으로 구성요소가 가로로 정열된 모양으로 놓여진 것을 나타낸다.
4)구성요소의 부 프레임명을 기재한 것으로 회사명 그룹에서 마크항목에 대하여 기술한다는 의미이다.
5. 구성요소의 위치관계를 나타내는 것으로 그 위치가 왼쪽 끝의 가장자리 위에 놓여진 것을 나타낸다.

이상의 방법으로 그림 1의 명함을 모델로하여 지식을 작성하여 지식 베이스로 하였고, 추가로 그림 12의 세로 쓰기모델을 지식 베이스로 하였다.

V. 명함화상의 구성소요추출

명함화상의 구성요소추출을 하려면 전처리, 문자레벨(level), 항목레벨, 그룹레벨등의 순서로 행한다. 문자의 블럭화는 Down-up블럭화 알고리즘[15]을 적용하였으며 항목후보 구형과 그룹후보구형의 추출은 다음과 같은 알고리즘을 제안하여 행하며, 각각의 후보구형들을 추출한다. 이렇게 추출된 각단의 후보구형을 앞장에서 정의하여 지식으로 표현 기술된 프레임의 지식 베이스와 매칭을 하여 각각을 검증하며, 매칭이 성공하면 그 구성요소는 결과로 추출한다.

본 논문에 전체 시스템의 흐름도는 그림 6과 같다.

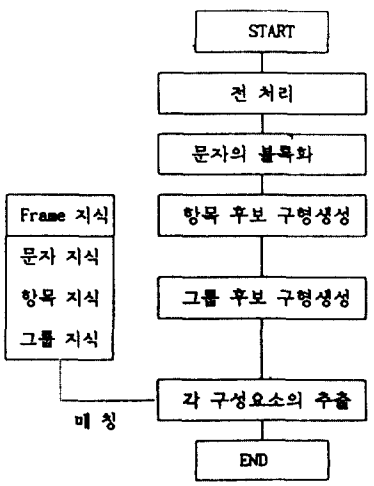


그림 6. 구성요소의 추출 알고리즘
Fig. 6. Extaction algorithm of component.

5-1. 문자 후보구형의 추출

각각의 문자들은 하나의 구형으로 구성되어 있으므로 문자의 위치와 크기등을 알기위하여 블록화를 행하여 문자 후보구형을 추출한다.

1)전처리

입력된 명함화상의 잡음이 문자로 처리되는 것을 방지하기 위하여 3*3 마스크를 씌워 고립점을 제거하였다.[16]

2)문자의 블록화

블록화는 오인권의 Down-up블록화 알고리즘을 적용하여 입력된 명함화상 데이타에서 우선 문자열을 파악한 다음, 한행에서 문자가 있는 부분의 최소값과 최대값을 구하여 배열에 저장하고 다음 행으로 옮겨가 마지막 행까지 그 작업을 반복 수행한다. 그런다음 다시 밑쪽에서 위쪽으로 스캔하면서 문자열을 인지하여, 그 문자열에 해당되는 문자의 최소값과 최대값 사이를 문자가 있는 부분이라 생각하여 블록화하게 된다. 이 과정에서 하나 이상의 문자가 붙어 하나의 블록으로 되는 경우와 하나의 문자가 상하 또는 좌우로 분리되어 나타나는 경우가 있게 된다. 전자의 경우에는 명함화상의 같은 행에서 문자분리점 추정이 거의 일정한 점과 블록 ratio(문자의 면적) 추적 및 histogram조사에 의한 방법에 의하여 분리하였다.

후자의 경우에는 한글, 영문, 특수문자에 공통적으로 하나의 문장열에서 위아래로 분리된 블록은 하나의 블록으로 간주하며 블록ratio 추적으로 통합하였다. 이 결과는 그림 7에 문자 후보구형이 생성되며, 문자 후보구형에서 얻어진 정보로부터 문자요소를 추출한다.

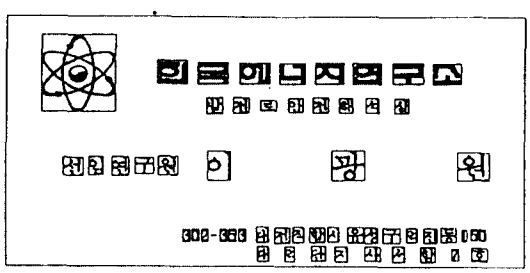


그림 7. 문자 후보구형
Fig. 7. Candidate rectangle of characters.

5-2. 항목 후보구형의 추출

5-1에서 얻어진 블록화된 문자 후보구형에서 항목 요소의 후보구형을 블록화하는 알고리즘은 다음과 같다.

1)블록의 가로폭(Xe－Xb)이 2이하이면 skip한다.

2)블록의 가로방향 시작점(Xb)을 다음 블록의 시작점(X_{b+1})과 비교하여 작은점을 찾는다.

3)블록의 세로방향의 시작점(Yb)을 다음 블록의 시작점(Y_{b+1})과 비교하여 작은점을 찾는다.

4)블록의 가로방향 끝점(Xe)을 다음 블록의 끝점(X_{e+1})과 비교하여 큰점을 찾는다.

5)블록의 세로방향 끝점(Ye)을 다음 블록의 끝점(Y_{e+1})과 비교하여 큰점을 찾는다.

6)만약 블록의 세로방향의 끝점(Ye)보다 다음 블록의 세로방향의 시작점이(Y_{e+1})이 크다면 항목 후보구형을 생성한다.

7)만약 블록의 세로방향의 시작점(Yb) 및 끝점(Ye)이 다음 블록의 세로방향에 시작점(Y_{b+1}) 및 끝점(Y_{e+1})의 차이가 6이상이고 블록의 가로방향의 끝점(Xe)과 다음 블록의 가로방향의 시작점(Y_{b+1})과의 차이가 11이상이고 다음 블록의 가로폭(Wx)이 21이상이면 항목구형을 생성한다.

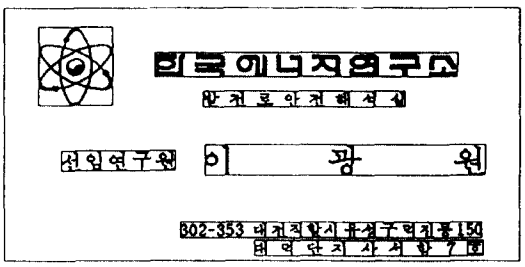


그림 8. 항목후보 구형
Fig. 8. Candidate rectangle of items.

표 4. 명함화상의 항목구형추출 지식
Table 4. Knowledge of item rectangle extract from name-card.

I1	마크는 1개만 존재한다.
I2	마크의 정방형도는 0에 가깝다.
I3	마크는 명함화상의 좌측상부에 위치한다.
I4	마크의 크기는 명함화상 중에서 가장 큰 문자나 기호로 되어있다.
I5	회사명은 1개만이 존재한다.
I6	회사명은 문자를 2자이상 15자이하로 구성되어 있다.
I7	회사명이 영문자로 재표현될 때는 회사명아래에 크기가 같거나 작은 크기로 문자수가 1.5배이상으로 구성되어 있다.
I8	회사명은 보통 명함화상의 중앙상부에 위치한다.
I9	회사명의 중심은 마크의 중심과 비슷하게 정열되어 있다.
I10	회사명은 문자의 크기가 가장크거나 이름의 문자크기와 비슷하다.
I11	소속부서명은 회사명의 아래에 위치한다.
I12	소속부서명은 이름의 위나 밑에 위치한다.
I13	소속부서명은 1개만이 존재한다.
I14	소속부서명은 문자의 크기가 회사명문자보다는 반드시 작다.
I15	소속부서명의 구성문자는 3자이상 15자이하로 되어있다.
I16	직함은 이름의 위에나 앞에 위치한다.
I17	직함은 문자 2자이상 5자이하로 구성된다.
I18	직함이 차지하는 면적은 이름이 차지하는 면적보다는 작다.
I19	직함의 문자크기는 이름의 문자크기보다는 작다.
I20	이름은 1개만이 존재한다.
I21	이름은 문자 2자이상 4자이하로 구성된다.
I22	이름의 문자크기는 직함의 문자보다 크고 회사명문자와 비슷하다.
I23	이름은 명함화상의 좌우상하 중앙부분에 위치한다.
I24	이름의 문자간격은 문자의 가로폭보다 1.5배이상 크다.
I25	우편번호는 이름아래와 주소사이에 있다.
I26	우편번호는 7개의 숫자 및 기호로 되어있다.
I27	우편번호는 문자크기가 이름문자의 1/2정도로 작다.
I28	주소는 명함화상의 아래부분에 위치한다.
I29	주소의 문자크기는 이름의 1/2정도의 작은문자로 되어있다.
I30	주소의 문자수는 10개이상 30개이하로 되어있다.
I31	주소는 가로로 최대 2줄까지 표현된다.

8)위 단계 1)-7)을 블록의 끝까지 반복 수행한다.

이상과 같이하여 생성된 항목 후보구형은 그림 8과 같고, 이들의 정보로부터 표4의 지식을 적용하여 기술된 프레임 지식 베이스를 매칭하여서 항목구형을 추출한다.

5-3. 그룹후보구형의 추출

5-2에서 얻어진 블록화된 항목후보구형 정보로부터 그룹요소의 후보구형을 블록화하는 알고리즘은 다음과 같다.

1)항목 후보구형의 중심 C(Cx, Cy)와 다음 항목 후보구형의 중심 C(C_{x+1}, C_{y+1})가 가로로 정열되어 있거나 그 차이가 10미만이면, 두 후보구형이 한 그룹이 된다.

이때의 가로방향의 시작점(Xb)의 작은점과 가장 큰 끝점(Xe)을 찾고, 세로방향의 가장 작은 시작점(Yb)과 가장 큰 끝점(Ye)을 찾는다.

2)다음 후보구형이 상하관계로 위치할 때, 세로방향의 끝점(Ye)과 다음 후보 구형의 시작점(Y_{b+1})의 차와 그 다음 후보구형과의 차이를 비교한다.

그 차가 작으면 두 후보구형은 한 그룹이므로 가로방향에서 가장 작은 시작점(Xb)과 가장 큰 끝점(Xe)을 찾고, 세로방향의 가장 작은 시작점(Yb)과 가장 큰 끝점(Ye)을 찾는다.

3)위에 단계 1)-2)에 의해 그룹 후보구형을 생성하며 단계 1)-3)을 끝까지 반복수행한다.

본 논문에서 사용한 가로쓰기 명함의 그룹은 상하로 나란히 존재한다. 따라서 후보 구형 생성은 다음의 지식중에 G1과 G2를 이용하여 3개의 그룹 후보구형을 그림 9와 같이 얻었다. 이 3개의 그룹 후보구형중에서 다음의 G3지식에 의해서 그룹의 구성요소를 추출한다.

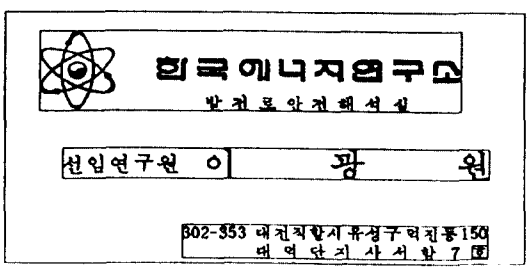


그림 9. 그룹 후보구형

Fig. 9. Cadidate rectangle of groups.

G1 ; 각 그룹은 1개씩 존재한다.

G2 ; 그룹내의 항목간격은 그룹간격보다 작다

G3 ; 각 그룹은 위에서 아래로 순서대로 회사명그룹, 이름그룹, 주소그룹의 순으로 존재한다.

Ⅵ. 실험 및 고찰

6-1. 실험환경 및 결과

실험대상으로는 국내에서 사용하고 있는 학계 및 전산관련 직종의 가로쓰기명함 100개를 표본으로 하였다. 명함화상 데이타는 영상 스캐너를 이용하여 640＊400 비트의 크기로 2치화한 것을 쓴다. 실험은 문자, 항목, 그룹의 순으로 하고 항목추출에는 문자추출 결과를 그룹추출에는 항목 추출결과를 이용하였다. 단, 추출은 이름그룹과 회사명그룹에서 항목을 추출하였다.

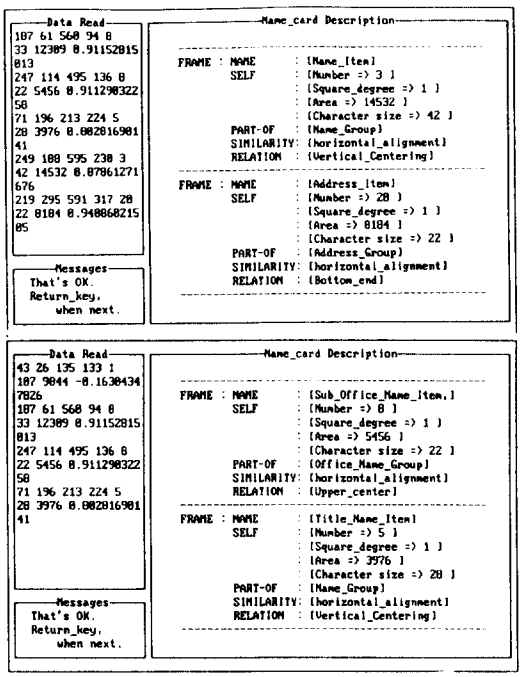
실험장비는 SymBios(PC/AT)시스템으로 잡음(noise)제거, 기본구형 정보추출등은 C언어를 사용하였고, 구성요소분류는 5절의 각 레벨의 특징을 이용하고 그림 5의 프레임전개를 사용하여 지식 베이스로 갖고, Turbo-PROLOG언어를 사용하여서 매칭처리를 행하여서 추출을 하였다. 특히 본 실험에서는 지식 베이스로 그림 8과 같은 형태를 가진 학교와 관공서의 명함을 지식 베이스로 갖고 있다.

Prolog로 매칭처리 한 결과는 그림 10(a)로서 그림 상단의 박스에는 회사 마크와 회사명(8자로 구성된 "한국에너지연구소")이 분류되었는데 이는 회사명 그룹에 속한다. 또 하단의 박스는 부회사명(8자로 구성된 "발전로안전해석실")과 직함(5자로 구성된 "선임연구원")으로 분류 되었다. 여기서 추출된 각 항목정보는 이미지로 그림 10(b)과 같이 각 항목별로 최종 출력되어 이를 항목 데이타 베이스에 저장한다. 이들 실험결과는 100개의 표본에 대한 통계로 표 5와 같이 얻었다.

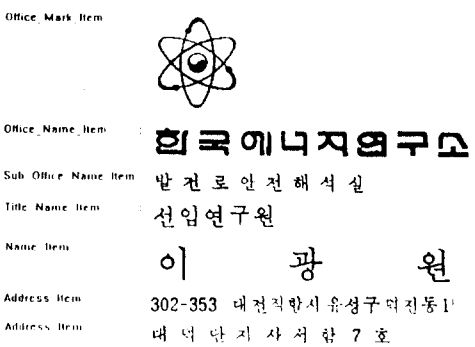
표 5. 실험결과

Table 5. Test result

대 상		후보구형생성(%)	구성요소추출(%)
군		100	100
항목	마 크	95	95
	회 사 명	97	97
	소속부서	90	90
	직 함	88	88
	이 름	100	100
평 균		95	95



(a)매칭 결과



(b)데이타 베이스를 위한 결과

그림 10. 그림 8의 추출결과의 부분
Fig. 10. Part of extraced result by fig. 8.

6-2. 고 찰

명함화상의 구성요소추출을 위해서는 정확한 문자구형 생성이 매우 중요한데 명함의 구성문자는 실제로 다양한 문자의 종류와 서체 및 크기로 이루워져 있으므로 이로 인한 오류가 있었다. 실험결과에서 보여 주듯이 항목후보구형으로 회사명과 이름을 생성한 경우에는 구성요소의 추출이 정확했다. 그러나 마크부분에 있어서의 생성과 추출율은 마크가 없는 경우에 회사명은 정확히 추출되나 마크의 크기가 회사명의 크기와 같은 경우에 오류 예 그림 11(a)와 같이 마크부분과 회사명부분의 분류가 잘 안된 경우였고, 소속부서 항목에서도 오류 예 그림 11(b)와 같이 소속부서와 직함이 구분이 어려운 경우가 있어 추출율이 낮게 나타났다.

주소그룹의 경우는 매우 작은 크기의 문자와 문자간의 밀접한 간격등으로 스캐너의 정밀도에 따라 정확한 후보구형생성에 문제점이 있었다.

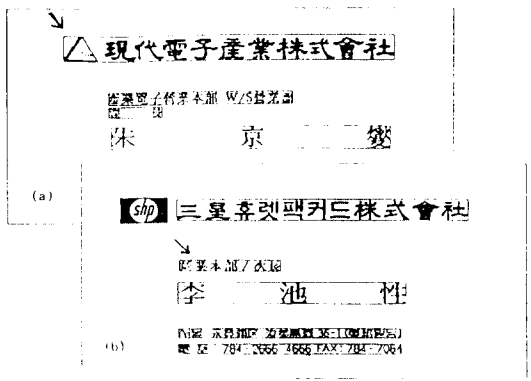


그림 11. 오류 예
Fig. 11. Example of ill-classification.

한편 본 시스템에다 세로쓰기형태의 명함화상을 처리하여 보면 그림 12와 같이 가로쓰기와 거의 비슷한 동작을 보였으나 마크부분은 추출이 않되었으며, 각 문자열이 가로로 누워진 형태의 결과를 보였다. 따라서 본 시스템을 이용한 세로쓰기형태의 명함 화상처리에서는 시스템의 지식베이스를 확장하므로서 처리 가능함을 보여주었다.

본 실험에서 1장의 명함을 스캐너를 이용하여 화상데이터로 받는데 약 3분 40초의 시간과 각 구성요소를 추출하는데 약 30초의 시간이 소요되었다. 앞으로 정밀한 화상데이타의 입력과 예외적인 경우에 대한 후보구형 생성규칙(rule)의 확충 및 정확한 구성요소 추출을 위한 지식원의 보완이 필요하다.

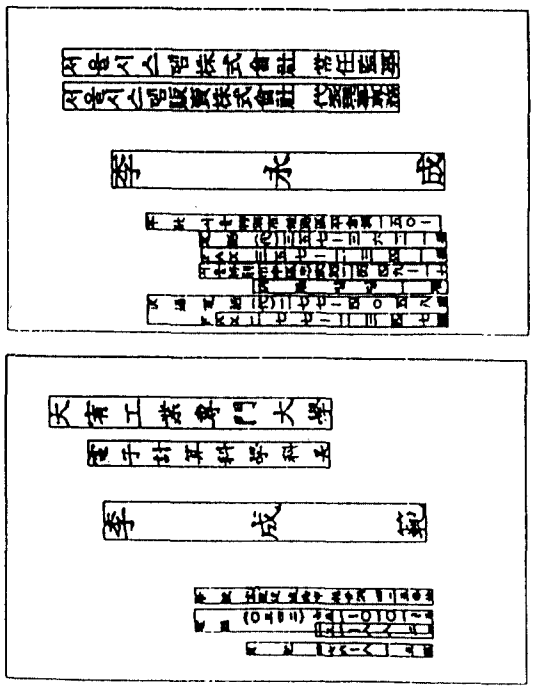


그림 12. 세로쓰기 예
Fig. 12. Example of vertical writing.

Ⅶ. 결 론

본 논문에서는 지식베이스를 이용하여 명함의 구성요소추출에 대하여 연구하였다. 이를 위하여 계층적인 특성을 가진 명함을 지식으로 갖기위한 기술방법이 컴퓨터에서는 중요한데, 제안한 방법에서는 명시적이고 기술이 용이하도록 프레임 술어를 사용하여 그 유용성이 실험으로 입증되었다. 구성요소추출을 위해, 각 영역의 블록화 알고리즘을 제안하여서 항목후보구형과 그룹후보구형을 생성하였다. 종래의 방법은 신문영상을 약 13분 정도 소요되었으나, 본 알고리즘을 써서, 명함의 항목을 블록화하는 데 약 10초가 소요되었다. 이들의 정보를 지식베이스에 저장된 지식모델에 매칭 시키므로서 명함의 구성요소가 변화하여도 유연하게 대응하여 구성요소를 이미지로 추출할 수 있었으며, 실험대상으로 설정한 국내 학계 및 전산관련직종의 명함 100개중에서 회사명그룹과 이름그룹의 항목추출율이 95%로 충실히 추출되었다.

본 논문의 실험결과는 명함의 정보를 자동으로 인식하기 위한 전처리 단계로, 항목부분을 이미지로 추출하므로서, 명함정보의 자동 데이타 베이스화의 가능성을 발견하였으며, 실험을 통하여 본 논문의 유용성을 확인하였다. 앞으로 한자, 한글 및 특수문자를 인식하는 툴이 개발되면, 본 연구에서 추출된 항목을 코드화 할 수 있게 되어, 완벽한 명함의 장동 인식시스템이 실현될 것이다. 문제점으로는 개인적으로 다양한 변화가 있는 명함에서는 많은 지식 베이스를 갖고 있거나, 학습할 수 있어야 한다는 것이다. 그러나 제한되고 거의 규격화된 문서에서는 유용함이 입증되었다.

참 고 문 헌

1. Wang D. and Srihari S. N., "Classification of Newspaper Image Blocks Using Texture Analysis," Computer Vision, Graphics, and Image Processing 47, 327-352, 1989.
2. Tang Y. Y., Suen C. Y., Yan C. D. and Cheriet M., "Document analysis and understanding : a brief survey." First International Conference on Document Analysis and recognition.(ICDAR), September 30-October 2, Saint-Malo, France, Vol.1, pp.17-31, 1991.
3. Kerpedjiev S. M., "Automatic extraction of information structures from documents," ICDAR, Saint-Malo, France, Vol.1, pp.32-40, 1991.
4. Kreich J., Luhn A. and Maderlechner G., "An experimental environment for model based document analysis," ICDAR, Saint-Malo, France, Vol.1, pp.50-58, 1991.
5. Chenevoy Y. and Belaid A., "Hypothesis management for structured document recognition," ICDAR, Saint-Malo, France, Vol.1, pp.121-129, 1991.
6. 김형훈, 이성환, 김진형, "한국 신문 영상의 구조분석을 통한 기사의 추출," 한국정보과학회논문지, 제15권 제5호, pp.392-404, 1988.
7. 이승형, "문서인식을 위한 한글과 한자의 구별과 한글의 형식분류에 관한 연구," 광운대학교 대학원 석사학위 논문, 1990.
8. Kise K., Babaguchi N. and Tezuka Y., "A proposal of reasoning and control for document im-

age understanding," EIC, PRU 89-76, 1989.
9. Luo Q., Watanabe T., Yoshida Y., Inagaki Y. and Saito T., "Understanding of Library Cataloging Cards on the Basis of a Knowledge-based Approach" 정보처리학회논문지, Vol.31 No. 12, pp.1755-1767, 1990.
10. Yeh P. S., Antoy S. A., A. Litcher and A. Resenfeld, "Address location on envelopes," Pattern Recognition, 20,2, pp.213-227, 1987.
11. Niyogi D. and Srihari S. N., "A rule-based system for document understanding," Proc. of AAAI-86, pp.789-793, 1986.
12. 가록현, 김신용, 박세진, 정동석, "CRLCA(Counted Run Length Cutting Algorithm)을 이용한 문서 분할에 관한 연구," 대한전자공학회 추계종합학술대회 논문집, 제15권, 제2호, pp.492-496, 1992.
13. Kise K., Sugiyama J., Babaguchi N. and Tezuka Y., "Layout model based analysis of document structure," EIC, PRU 89-75, 1989.
14. J.H. Connell and M. Brady, "Generating and Generalizing Models of Visual Objects," Artificial Intelligence, 31, pp.159-183, 1987.
15. 오인권, "영문이 혼합된 한글문서에서의 문자 및 특수문자추출에 관한 연구," 광운대학교 대학원 석사학위 논문, 1988.
16. 남궁재찬, "화상공학의 기초," 기전연구사, pp.131-133, 1989.

李 成 範(Sung Bum Lee) 정회원
1973년 : 한양대학교 전기과공학과 졸업(공학사)
1981년 : 동국대학교 대학원 전기공학과 졸업(공학석사)
1989년 : 광운대학교 대학원 전자계산기공학과 재학
1981년~현재 : 대유공업전문대학 전기과 부교수
※관심분야 : 패턴인식, 인공지능, 컴퓨터 비젼

南宮在贊(Jae Chan NamKung) 정회원
1970년 : 인하대학교 전기공학과 졸업(공학사)
1976년 : 인하대학교 대학원 전자공학과 졸업(공학석사)
1982년 : 인하대학교 대학원 전자공학과 졸업(공학박사)
1982년~1984년 : 일본 TOHOKU대학 객원 교수
1979년~현재 : 광운대학교 전자계산기공학과 교수
※관심분야 : 패턴인식, 컴퓨터 비젼, 인공지능