획 추출에 의한 한글 문서 인식 시스템의 설계 및 구현

이 관 용 , 이 일 병
연세대학교 전산과학과

Design and Implementation of Hangul Document Recognition System by Stroke Extraction

Kwanyong Lee and Yillbyung Lee
Department of Computer Science, Yonsei University

요 약

입력된 문서 영상으로 부터 분리 추출된 문자 영상을 올바르게 인식하는 것은 문서 인식에서 가장 핵심적인 부분이다. 스캐너를 통해 입력되고 분리된 실제의 문자 영상은 많은 문제점들을 가지고 있다. 한글의 경우 이 중 개별 문자 영상내의 각 자소간의 접촉은 올바른 인식을 저해하는 주요한 원인이다. 이런 접촉의 문제를 효율적으로 해결하기 위해 한글의 구조적 특성을 지닌 "방향 필터"를 정의하고, 이것을 이용하여 세선화된 문자 영상을 추적하면서 선소들을 뽑아낸다. 이렇게 하여 얻은 선소들과 선소들간의 지식을 조합하여 한글 자소 획을 추출케 되고 결국에는 이런 획의 조합을 통해 문자 영상을 인식하는 방법을 제안한다.

Ⅰ. 서론

컴퓨터를 우리의 생활과 동떨어져 생각할 수 없게 된 현 정보화 사회에 있어서 대부분의 데이타 혹은 정보가 인간의 수고와 노력을 요구하는 기존의 입력 방법인 자판기(keyboard)를 통해 입력이 이루어져 왔다. 하지만 산업이 점점 다양화되고 전문화되어 가면서 그에 따른 정보의 양 또한 급속히 증가할 것이다. 따라서 기존의 입력 방법을 고수한다면 막대한 데이타가 보다 신속 정확하게 처리되지 못할 뿐 아니라 전산 입력과 관리가 용이하지 못할 것은 자명한 사실이다. 이러한 사실로 부터 영상 입력 장치를 통한 문서의 자동 입력에 대한 연구의 필요성이 일찌기 대두되었고 영문의 경우에는 인쇄체 영문서나 제한된 필기체를 인식하는 시스템까지 실용화되었고, 한글 문자의 인식에 대한 시도는 1960년대 말에 시작되어 그 연구가 계속되어지고 있다. 하지만 한글 문서의 자동 입력에 대한 연구

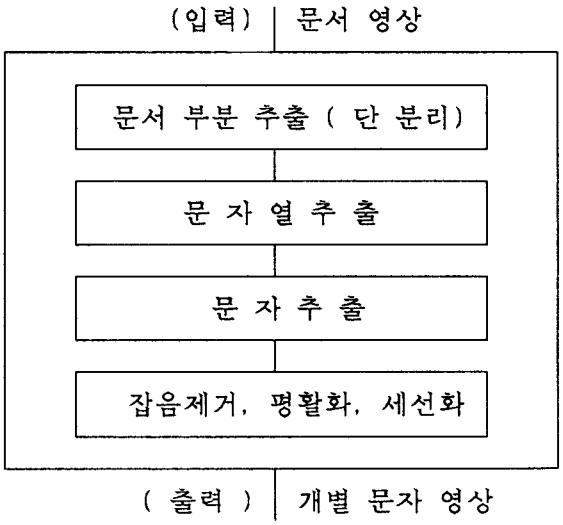
본 연구는 1989년도 연세대학교 학술연구비 (이일병:"사전편찬을 위한 한글문서 인식시스템 구현 연구")에서 지원 받았음을 밝힙니다.

가 시작된 것은 얼마되지 않았고, 아직도 그 연구의 방법 및 결과가 많은 문제점을 지니고 있어 실용화되기에는 보다 많은 연구가 필요한 형편이다.

문자 영상에서 각 자소간의 접촉으로 인하여 그 자소가 다른 자소로 인식되거나 다른 자소의 획을 침범하여 오인식이 되는 것을 해결하기 위한 연구가 있었다[1]. 하지만 자소간 접촉의 문제에 대한 좀 더 보편 타당한 방법이 제시되지 못하여 문서 인식 시스템 실용화에 큰 장애가 되어왔다. 본 연구에서는 기존의 인식 방법이 이러한 문제에 대해 가지는 취약점을 보완하기 위해 문자 영상에 대해 시간적 정보와 각 자소의 위치 정보등을 구해내어, 온라인 인식과 유사한 방법으로 시작점과 끝점과 방향값을 계산하여 인식하는 방법을 사용한다. 하지만 기존의 온라인 인식 방법은 획을 단위로 추출 인식하는[2,3,4] 반면 본 방법은 획보다 작은 단위인 선소를 인식 단위로 생각하고 그 선소의 조합으로 획의 추출을 가능케 한다. 선소를 추출하기 위해 한글의 구조적 특성을 지닌 "방향 필터"를 정의하고 그것을 이용하여 시간적 정보와 위치 정보를 얻어내고, 이런 정보을 이용하여 자소간의 접촉을 효율적으로 해결하면서 문자의 인식을 수행하게 된다.

Ⅱ. 전처리 단계

전처리 단계는 입력 문서 영상에 대해서 문자 영역과 비문자 영역을 분리한 다음 문자 영역에 대해서 개별 문자를 분리 추출하여 인식 과정에서 처리될 수 있도록 원하는 형태의 데이타로 문자 영상을 개선 및 변환시켜 주는 과정이다. 하지만 이 부분에 대한 연구[5,6,7]는 아직도 미약하고, 연구된 것들 조차도 아직은 처리 방법 및 속도 또한 그 결과에 대해서 많은 문제점을 가지고 있다. 따라서 실제 모든 가능한 문서에 대해 전처리 단계의 알고리즘을 적용하는 것은 아직도 많은 연구가 되어야 할 부분이다. 본 연구에서는 가능한 문서에 대한 제한점을 가하여 실험하고 하고 있다. 전처리 단계의 전체적 개요는 아래 그림과 같다.



< 그림 1 > 전처리 단계의 개요

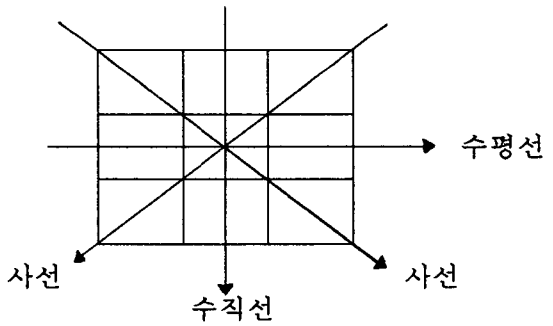
문서 영상에 대해서 가로 문서인지 세로 문서인지 구별하기 위해 가로 세로에 대한 투영 (projection) 화일을 이용하여 문서 부분(segment)을 추출한 후 문서 형태를 구분하고, 문자열을 추출한다. 이렇게 얻은 각 문자열에 대해서 역시 동일한 투영 방법을 사용하여 개별 문자를 추출하게 된다. 하지만 이 과정에서 문자간의 접촉이 발견되며 이러한 접촉은 몇 가지 유형으로 분리된다[6]. 접촉된 문자들은 글자 크기의 정보를 이용하여 강압적으로 분리된다. 이러한 개별 문자에 대해 잡음제거 , 평활화, 선형화, 세선화 과정을 거쳐서 얻은 결과가 전처리의 출력으로 인식 과정에 전달되어 인식이 수행된다.

Ⅲ.문자 인식 단계

Ⅲ-1. 방향 필터

한글은 구조적으로 기본 자소 24자의 조합이고, 이러한 자소는 획의 조합으로 간주된다.[8] 그리고 획은 기본적으로 수평선 , 수직선, 사선, 곡선등의 더 작은 단위인 선소로 구성된다. 특히 한글의 경우에는 영문자와 달리 곡선의 선소보다는 수평선, 수직선의 선소가 주를 이루고 약간의 사선이 추가된 형태로서 획이 형성된다고 볼 수 있다. 따라서 인쇄 문자에서는 획을 하나의 단위로서 직접 추출하여 인식하기 보다는 선소 단위의 추출이 훨씬 용이하고, 이렇게 검출된 선소의 조합을 통해 획을 인식하는 방법이 효과적이라고 보인다.

방향 필터란 3x3 윈도우로서 각 위치에 고정된 가중치를 가지고 있어 수평선, 수직선, 사선의 선소를 추출케 해 주는 윈도우이다. 아래의 그림은 3x3 윈도우에서의 각 선소를 표시한다.



< 그림 2 > 3 x 3 윈도우에서의 각 선소에 대한 표현

방향 필터의 각 위치에 대한 가중치를 얻기 위해 우선 다음과 같은 가정을 하자.

P_1	P_2	P_3
P_8	(XY)	P_4
P_7	P_6	P_5

W_1	W_2	W_3
W_8	(XY)	W_4
W_7	W_6	W_5

$$\begin{cases} (X, Y) : \text{현재의 좌표} \\ P_i = \begin{cases} 1 & , \text{만약 윈도우 위치 } i \text{ 가 글자 영역이면} \\ 0 & , \text{그렇지 않으면} \end{cases} \\ W_i : \text{윈도우 위치 } i \text{ 의 가중치} \end{cases}$$

$$C_n \equiv \left\{ \sum_{i=1}^{8} P_i = n \right\}$$

C_n은 좌표 (X, Y)에서 연결된 점의 수를 나타내며 본 연구에서는 불완전한 8 방향 연결성을 가지는 세선화 알고리즘을 사용하였으므로 C_1, C_2, C_3, C_4가 정의되어 사용되었다.

$$W_{xy} \equiv \sum_{i=1}^{8} W_i * P_i$$

W_{xy}는 좌표 (X, Y)에서 방향 필터 가중치의 합이다.

방향 필터의 가중치들은 아래의 조건을 만족하는 W_i ($i = 1, 2, \cdots, 8$)로서 구성된다.

① $i, j \in \{1, 2, \cdots, 8\}, \quad i \neq j \rightarrow W_i \neq W_j$

② $\begin{cases} W_2 = -(W_4) \\ W_8 = -(W_6) \end{cases}$ or $\begin{cases} W_2 = -(W_8) \\ W_8 = -(W_4) \end{cases}$

③ $\text{Max}\{W_1, W_3, W_5, W_7\} < \text{Min}\{|W_2|, |W_4|, |W_6|, |W_8|\}$

④ $\forall a_{xy}, a_{x'y'} \in C_n \rightarrow W_{xy} \neq W_{x'y'}$
for each n = 1, 2, 3, 4
($x \neq x'$ and $y \neq y'$)

이렇게 하여 정의된 방향 필터에 대해서 각 선소들은 다음과 같은 범위의 방향값을 가짐을 알 수 있다.

① $\text{Min}\{ W_1, W_3, W_5, W_7 \} \leq$ 사선 $\leq \text{Max}\{ W_1+W_5, W_3+W_7 \}$

② if $W_2 < 0$

$\text{Min}\{ W_1+W_4, W_4+W_7, W_3+W_8, W_5+W_8 \} \leq$ 수평선 $\leq (W_4+W_8)$

$(W_2+W_6) \leq$ 수직선 $\leq \text{Max}\{ W_2+W_7, W_2+W_5, W_3+W_6, W_1+W_6 \}$

else

$(W_2+W_6) \leq$ 수평선 $\leq \text{Max}\{ W_2+W_7, W_2+W_5, W_3+W_6, W_1+W_6 \}$

$\text{Min}\{ W_1+W_4, W_4+W_7, W_3+W_8, W_5+W_8 \} \leq$ 수직선 $\leq (W_4+W_8)$

현재의 좌표 (X, Y)를 이용하여 다음에 추적해야 할 좌표를 찾는 함수 $f(x,y)$ 가 필요

하다. 하지만 Cn (n = 1, 2, 3, 4)의 가지수가 많아질수록 함수 $f(x,y)$ 로서는 다음의 좌표를 효율적으로 찾을 수 없을 뿐더러 접촉이 발생하는 경우 다음에 추적해야 할 점을 결정하는 것이 힘들다. 따라서 현재의 좌표 (x, y)를 이용하는 대신에 연결된 점의 수 Cn과 현재 좌표에서의 가중치의 합 Wxy와 시간 t를 이용하는 함수 $N(C_n, W_{xy}, t)$를 정의하여 사용할 수 있다.

$$f(x, y) \equiv N (C_n, W_{xy}, t)$$

함수 N에 의해 다음으로 추적할 점을 결정하기 위해서는 반드시 Cn (n=1,2,3,4) 에 해당하는 경우를 미리 정의되어야 하고, 본 연구에서 사용된 Cn (n = 1, 2, 3, 4)의 갯수는 <표 1>과 같다.

Ⅲ-2. 자소의 처리 순서

현재까지 한글 문자 인식에서는 각 자소의 처리를 수평모음 또는 수직 모음을 먼저 처리하고 초성자음이나 종성자음을 처리하는 연구[9,10]와 모음의 우선 처리를 가정하지 않는 연구들이 주를 이루었다[11,12]. 하지만 본 연구에서는 인간이 개별 문자를 처리하는 방식이 글자를 쓰는 방식과 흡사하여 초성자음 , 수평모음, 수직모음 그리고 종성자음의 순으로 이루어진다고 생각하고 이 순서를 따라서 처리하면 접촉의 경우를 해결하는데 도움이 될 것으로 생각했다.

문자 영상이 들어 있는 2차원 배열의 좌상의 좌표에서 부터 윈도우를 탐색하면서 처음으로 방문되지 않은 채 발견되는 점을 시작점으로 택하고, 그 점에서 앞으로 이 선소가 어떤 방향으로 진행할 지를 결정하고 그 점에서 부터 추적을 시작한다. 하지만 윈도우를 탐색할 때 초성자음 , 수평모음, 수직모음 그리고 종성자음순으로 시작점을 찾기 위해서는 탐색시에 기울기의 조정이 필요하게 된다. 따라서 기울기의 조정을 용이하게 하기 위해 기울기를 ½과 2로 고정시켜 두 종류의 기울기로서 탐색을 시작하여 처음으로 방문되지 않은 점들을 각각 찾아낸다. 이렇게 찾아낸 두 점에 대해서 각 자소의 시작점에 대한 위치 정보를 고려하여 최종적으로 시작점을 결정한다. 각 자소의 시작점에 대한 위치 정보를 고려하기 위해서 <그림 3>과 같은 M x N 배열에서의 각 자소의 시작점이 존재할 영역을 정의하였다.

Ⅲ-3. 인식 방법

한 자소를 인식하기 위해서는 우선 그 자소의 시작점에서 결정된 선소의 방향과 방향값의 최소치와 최대치를 유지한다. 함수 $N(C_n,W_{xy},t)$를 이용하여 다음 점을 구해 추적하면서 필터 가중치의 합을 구한다. 이 합이 유지되고 있는 선소의 방향값의 최소치와 최대치 사이에 존재하면 현재의 좌표에서 선소의 방향에 변화가 발생하지 않은 것으로 다음 점을 구해 추적하면서 변화 여부를 계속 조사한다. 만약 필터 가중치의 합이 최소치와 최대치

사이에 존재하지 않으면 선소 방향값에 변화가 발생한 것으로 간주되어 현재까지 추적된 점들을 선소로서 추출한다. 이렇게 추출된 선소와 기존에 추출된 선소의 연결 관계, 위치 관계, 시간 정보등의 지식을 이용하여 획을 추출한다. 획의 조합을 통하여 자소를 인식하고 결국에는 자소들을 조합하여 문자를 인식하게 된다.

자소간 접촉의 대부분의 경우는 C_3 ,즉 분기점 형태로 발생한다. 따라서 분기점에서 접촉이 발생한 경우 그것을 처리하기 위해서 분기점에서의 가능한 변환 경우를 고려하기 보다는 현재점까지의 정보 (시간 정보, 방향정보, 자소에 대한 정보, 위치 정보)등을 참조하여 함수 $N(C_n,W_{xy},t)$이 우선 추적 방향을 선정하여 추적케한다. 이렇게 추적된 선소들이 만약 자소를 형성하지 못하게 되면 현재까지의 추적된 점들이 복구되고 우선 추적 방향이 수정되어서 다른 경우를 추적하게 된다.

문자 영상에서 사선 방향의 선소들은 대부분 수평선과 수직선이 계단형으로 구성되기 때문에 하나의 사선으로 추출되기 보다는 여러개의 나누어진 수평 선소와 수직 선소로 구성된다. 이러한 것들을 처리하기 위해서는 분리된 선소들을 하나의 선소로 합치어 주는 과정이 필요하게 된다.

위의 과정들을 종합하면 아래의 알고리즘과 같이 표현된다.

```
For each grapheme
 Find Start Point : (x,y)
 while ( Not recognize a grapheme )
   Compute Wxy , θmin , θmax
       /* θmin, θmax는 Wxy가 해당하는
          선소 방향값의 최소값, 최대값   */
   while ( θmin ≤ Wxy ≤ θmax )
     (x, y) ⇐ N(Cn,Wxy,t)
     Compute Wxy
   if ( ∃ splitted stroke ) Combine
   else Recognize Stroke-Grapheme
   if ( Buffer is not Empty ) Delete a point: (x,y)
```

Ⅳ. 실험 결과 및 결론

본 논문에서 제안한 방법의 타당성을 검토하기 위해 300 DPI의 해상도를 가지는 스캐너를 사용하여 한글 문서 데이타를 입력받아 실험에 사용하고 있다. 현재까지 진행된 실험을 통해 볼 때 C_3 형태의 접촉에 대해서는 다른 알고리즘보다 효율적으로 인식과정이 수행되었다. <그림 4>는 C_3 형태의 접촉에 대한 인식 결과 및 시간적 정보를 나타낸다. 하지만 C_2 형태의 접촉은 접촉점에서 별다른 특징이 발견되지 못하여 접촉을 허용된 자소가 오인식되는 결과를 보여 주었다. <그림 5>는 C_2 형태의 접촉으로 인하여 오인식되는 데이타이다. 그리고 'ㅎ'과 같은 것은 한 자소내의 접촉으로 인하여 세선화된 결과가 도저히 'ㅎ'으로 인식하기 어려운 형태로 나타나는 문제점을 가지고 있다. 앞으로 이러한 문제점들에 대해

보다 많은 연구가 진행되어야 하고 더 나아가 세선화된 문자 영상이 아닌 원래의 문자 영상을 추적하여 인식하는 방법도 함께 연구됨이 바람직하다.

참고 문헌

[1]. 강 현철, "구문 분석과 패턴분류를 이용한 한글 인식에 관한 연구", 연세대학교 박사학위 논문, 1989

[2]. 이 희동, 김 태균, "보강문맥 자유 문법을 이용한 필기체 한글 온라인 인식," 대한전자공학회 논문지 제16권5호 , pp.37-44, 1989

[3]. 강 윤주, 김 준호, 박 종화, "On-Line 한글 인식," 연세대학교 전산과학과 학사학위 논문, 1989

[4]. 정 아연, "동적 문자 인식을 위한 신경망 모형 Sejong-Net의 설계," 연세대학교 석사학위 논문, 1989

[5]. 김 태균 등, "한글 문서 인식을 위한 문서 영상에서의 문자와 그림의 분리 추출," 한글날 기념 학술 발표대회 논문집, pp 50 -53, 1989

[6]. 이 균하 등, "한글문서에서의 낱자 분리 알고리즘," 한글날 기념 학술 발표대회 논문집, pp 203-208, 1989.10

[7]. 남궁 재찬, 류 황빈, 남궁 윤, "한국어 문서로부터 문자분리 및 도형 추출에 관한 연구," 대한전자공학회 논문지, 제25권9호, pp.1091-1100, 1988.10

[8]. 이 주근, "한글 문자의 인식에 관한 연구 (IV)," 대한전자공학회 논문지, 제9권4호, pp. 197-204, 1972

[9]. 고 견 , "한글 문서 인식 시스템에 관한 연구," 연세대학교 석사학위 논문, 1988

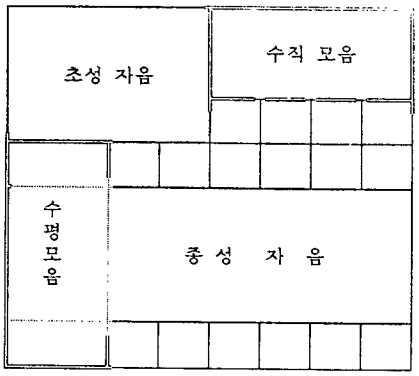
[10]. 이 주근 등, "한글 Pattern에서 Subpattern 분리와 인식에 관한 연구," 전자공학회지 , Vol.18, No. 3 pp 1-8, June. 1981

[11]. 이 승호, "구조적 한글 인식에 위한 자획 추출에 관한 연구," 한국과학기술원 전산학과 석사학위 논문, 1988

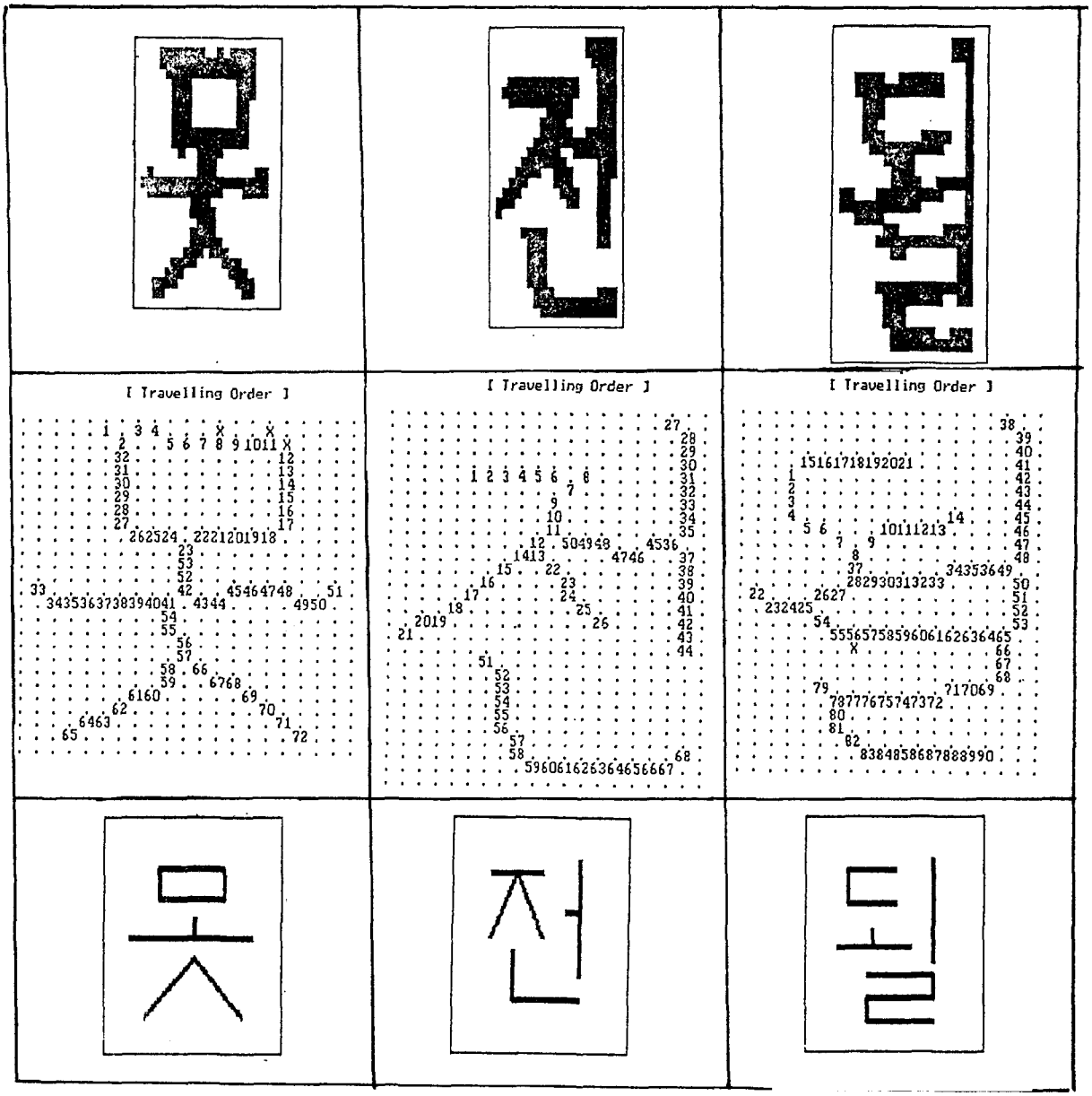
[12]. 하 진영, 김 진형, "학습 기능을 이용한 필기 한글 인식에 관한 연구," 춘계 인공지능 학술발표회 논문집, pp.3-24, 1989

C_n	갯 수
1	8
2	20
3	32
4	4

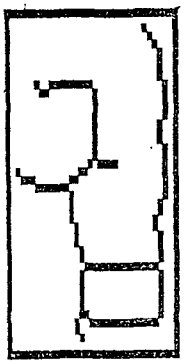
< 표 1 > 본 연구에 사용된 C_n의 종류와 그에 해당하는 갯수



< 그림 3 > M x N 배열에서 각 자소의 시작점이 존재하는 영역



<그림 4> C_3 형태의 접촉을 가진 데이타의 인식 결과 및 시간적 정보



<그림 5> C_2 형태의 접촉으로 인한 오인식된 예