주소 인식 시스템을 위한 필기 한글 단어 인식

권진욱 이관용 변혜란 이일병
연세대학교 컴퓨터과학과

Handwritten Korean Word Recognition for Address Recognition

Jinwook Kwon, Kwanyong Lee, Hyeran Byun, Yilbyung Lee
Dept. of Computer Science, Yonsei University

요 약

최근 주소를 자동으로 인식하여 우편물 분류와 같은 업무를 효과적으로 수행하기 위한 연구가 진행되고 있다. 기존 연구들은 낱자 단위의 인식을 수행한 후 사전 형태의 간단한 DB를 통해 최종의 결과를 생성한다. 그러나 한글과 같은 복잡한 구조의 필기 문자에 대한 인식기의 성능은 아직도 미흡한 상태이다. 따라서 낱자 인식기의 성능에 의존하는 현재와 같은 방법으로는 만족할 만한 결과를 얻기가 힘들 것으로 생각된다.

본 논문에서는 낱자 인식 결과에 크게 의존하지 않고 주소에 나타나는 단어의 낱자들 사이간 연결 정보를 이용하여 단어를 인식할 수 있는 시스템을 제안한다. 본 시스템은 통계적 인식기를 사용하여 낱자를 인식하는 부분과 낱자 인식 결과를 종합하여 단어 수준의 인식과정을 통해 최종의 결과를 생성하는 부분으로 구성된다. 통계적 인식기는 Nearest neighborhood 방법을 사용하여 간단한 형태로 구현하였다. 단어인식 모듈은 단어에서 모든 문자간의 관계를 표현할 수 있도록 HMM 모형을 사용하여 어휘정보 네트워크를 구성하고 이를 이용하여 주소에 나타나는 단어를 인식하도록 하였다.

PE92 한글 문자 데이터를 이용하여 실험을 수행한 결과, 통계적 인식기의 성능이 저조함에도 불구하고 HMM을 이용한 어휘정보 네트워크가 이를 보완함으로써 좋은 결과를 얻었다. 이러한 단어 인식 방법은 주소 이외의 다른 단어 집합에 대해서도 쉽게 적용될 수 있을 것으로 예상된다.

1. 서론

최근의 문자 인식에 관한 연구들이 여러 방면에 걸쳐서 진행되고 있고 한글 인식에 관련된 연구도 다양하게 진행되고 있다. 특히 인쇄체 한글에 대한 연구는 상당한 정도의 진전을 이루고 있지만 필기체 한글에 대한 연구는 아직까지 그리 만족할 만한 수준은 아니다. 왜냐하면 필기체 한글의 인식은 말 그대로 사람이 손으로 쓰기 때문에 그 변형이 매우 다양하다는 점이다. 그러나, 한글이 이런 특성을 가지고 있어서 인식에 어려움이 많지만 한 가지 주목할 점은 실제 쓰이는 낱자의 수는 많이 제한되어 있다는 점이다. 그래서 이런 제한을 이용하여 한글을 인식하려는 시도가 여러 방면에서 행해져 왔다. 따라서 본 논문에서는 제한된 수의 낱자와 그 낱자들로 이루어진 단어 집합을 인식하는 시스템을 구성했다.

보통 인식의 단위로서 사용하는 것은 하나의 낱자나 또는 한 문자를 자소 단위로 분리하는 방법을 많이 사용하고 있지만, 본 논문에서는 단어 수준의 인식방법을 사용하였다. 기존의 단어 인식 방법은 획의 출현순서를 관측열(Observation Sequence)로하여 인식하는 방법이 주로 사용되고 있다.[2][3][4] 이런 방법들은 분할과정이 불필요하다는 장점이 있지만 복잡한 HMM 네트워크를 필요로 한다. 그리고 획에 관련된 정보뿐만이 아니라 어휘 수준의 정보까지 함께 이용하여 인식하는 연구가 수행되고 있다.[2]

본 논문에서 쓰인 단어 수준의 인식이란 다중특징을 사용한 낱자 인식기로 단어의 각 낱자를 인식한다. 그리고 낱자 인식기의 결과와 어휘지식과 관련된 낱자간의 연결정보가 학습된 어휘정보 네트워크를 사용하여 단어를 인식한다.

다음 2장에서는 단어인식 시스템의 구성을 설명하고, 3장에서는 실험환경에 관해 설명하며, 4장에서는 실험결과를 보이고, 5장에서 결론을 내린다.

2. 단어인식 시스템의 구성

단어 인식 시스템의 흐름도는 그림1과 같다. 통계적 인식기는 입력으로 낱자영상 집합을 받아 낱자의 인식 결과를 만들게 된다. 그리고 이 결과를 단어 인식기에 주게 되고 단어 인식기는 어휘정보 네트워크와 낱자인식의 확률값을 이용하여 단어인식 후보를 생성한다.

2.1 낱자인식에 사용되는 특징

낱자 인식에는 간단한 통계적 인식기가 사용되는데 여기에 쓰이는 특징은 두 가지이다.

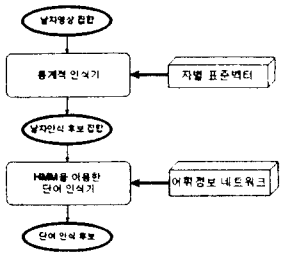

그림 1 단어인식 시스템 흐름도

첫째로 가로-세로 방향성분의 추출방법(그림2)이다.

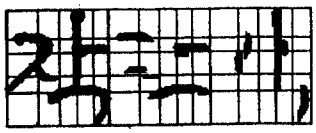

그림 2 가로세로 방향성분의 추출

이 특징을 사용하는 이유는 한글 글자체의 형태에 연유한다. 필기습관에 따른 변형이 존재하지만 한글은 기본적으로 가로획과 세로획이 대부분을 차지한다. 따라서 이런 가로획과 세로획의 성분을 추출하여 특징을 추출하는 것이 한글인식에 있어서는 타당한 방법이다. 그림2에서 볼 수 있듯이 한 글자를 가로방향 획과 세로방향 획을 분리한 후 동적 그물망을 사용하여 특징값을 구하였다.

둘째로는 Gradient 방향성분에 의한 특징을 사용하는데(그림3) 이 특징은 문자인식에서 방향성분을 나타내는 좋은 특징으로 알려져 있다.[7] 본 논문에서는 8방향 코드를 특징벡터로 사용하였다.

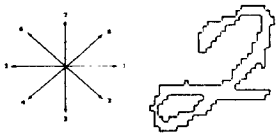

그림 3 Gradient 방향성분에 의한 특징

본 논문에서는 위에서 언급한 두 가지 특징의 크기를 정규화하여 낱자인식을 위한 특징으로 사용하였다.

2.2 통계적 인식기

2.1에서 언급한데로 특징이 추출되면 각 자별로 특징에 대한 표준벡터를 구하는데, 이번 실험에서 대상으로 하는 354자에 대한 표준벡터를 생성한 후 Nearest neighborhood방법을 사용하여 테스트 벡터와 각 낱자별 표준벡터와의 거리를 구한 후 낱자의 인식 결과로 사용한다. 다음 표1은 354개의 낱자를 인식해본 결과이다.

표1의 인식결과를 보면 정인식률은 54%이지만 후보를 80개까지 선정하면 99%정도의 인식률을 보여준다. 단어의 평균 길이를 4정도로 가정하면 $(0.99)^4$=0.96 정도이므로 단어 인식률은 최대 96%정도 얻을 수 있다. 따라서 단어인식의 입력으로서 낱자당 80개의 후보를 채용하였다.

후보의 개수	인 식 률
1	54
2	68
3	75
4	79
5	82
~	~
50	97
~	~
80	99

표 1 통계적 인식기의 후보 인식률

비록 여기서는 80개라는 많은 수의 후보를 채용하지만 단어 인식기에서 후보의 약 90% 정도가 부적절한 낱자로서 제거된다.

3.3 어휘정보 네트워크의 구성

본 논문에서 쓰이는 어휘정보 네트워크는 기존의 HMM모델을 이용하여 만든 것이다.(그림4)

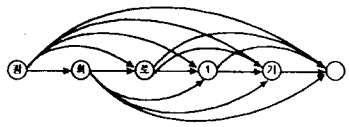

그림 4 어휘정보 네트워크의 예

그림4는 '광회로1가'라는 단어에 대한 연결모형이다. 그림에서 보는 것처럼 '광'이라는 낱자에 대해서 다른 모든 낱자에 대한 연결확률이 존재며 다른 나머지 낱자들도 마찬가지의 정보를 가진다. 그리고 네트워크의 길이는 단어의 최대길이인 6에 맞춰져 있으므로 위 그림의 마지막 노드값이 0이고 이와 연결된 모든 연결확률도 0이 된다.

위에서 사용된 모델의 특징은 인접하는 상태의 노드들 뿐만이 아니라 인접하지 않는 상태의 모든 노드들에 대해서도 연결확률이 존재한다. 이것은 네트워크가 단어에 있어서 인접하는 낱자들 사이의 연결확률뿐만이 아니라 인접하지 않는 낱자들 사이의 연결확률을 포함하는 것을 의미하며 부적절한 단어의 제거 및 후보단어의 인식 점수를 계산하는데 사용되어 진다.

어휘정보 네트워크의 확률값은 두 단계를 거쳐서 결정된다. 첫 번째 단계로는 발생빈도에 따른 조건부 확률에 따라 확률값을 일차적으로 결정하는 것이다.

$$P(SecondChar|FirstChar) = \frac{P(FirstChar \cap SecondChar)}{P(FirstChar)}$$

두 번째 단계로는 각 노드에서 다른 노드로 연결되는 연결확률의 값들의 평균과 표준편차가 같도록 조정하는 것이다.

3.4 어휘정보 네트워크를 이용한 단어 인식기

통계적 인식기에서 나오는 단어 인식의 결과는 단어의 각 위치별로 모두 80개씩의 후보로 구성되어 있다.

이 결과와 어휘정보 네트워크에 기억되어 있는 연결확률에 따라 Viterbi 알고리즘을 변형시켜 적용한다. 단어 인식 알고리즘은 다음과 같다.

1. **남자영상에서 특징을 추출한다.**
2. **통계적 인식기를 사용하여 인식한 후 상위80개를 후보로 선정한다.**
3. **인접해있는 상태(state)에 있는 노드들과 연결을 가지지 않는 노드를 제거한다.**

$$C_k : k_{th}\ char\ candidate\ set$$
$$i \in C_k,\ j \in C_{k+1}$$
$$If \sum_{\forall j} H_{ij} = 0,\ then\ delete\ i$$

4. **인접하지 않는 상태(state)에 있는 노드들과 연결이 존재하지 않는 노드를 제거한다.**

$$C_k : k_{th}\ char\ candidate\ set$$
$$i \in C_k,\ j \in C_{k+2}$$
$$If \sum_{\forall j} H_{ij} = 0,\ then\ delete\ i$$

5. **남아있는 남자 후보들로 가능한 모든 경로를 조사하며 인접하는 남자간의 연결이 존재하지 않는 경로를 삭제한다.**

$$Path\ set\ P = \{p_1, p_2, \ldots\ldots, p_n\}$$
$$p_i = \{C_1, C_2, \ldots\ldots, C_N\}$$
$$For\ each\ p_i,$$
$$If \prod_{j=1}^{N-1} H_{C_jC_{j+1}} = 0,\ then\ delete\ p_i$$

6. **5에서 남은 경로중 인접하지 않는 남자간의 연결을 모두 조사하여 연결이 하나라도 존재하지 않으면 제거한다.**

$$Path\ set\ P = \{p_1, p_2, \ldots\ldots, p_n\}$$
$$p_i = \{C_1, C_2, \ldots\ldots, C_N\}$$
$$For\ each\ p_i,$$
$$If \prod_{j=1}^{N-2} H_{C_jC_{j+2}} = 0,\ then\ delete\ p_i$$

7. **남아있는 경로들의 인식점수를 계산하고 순위를 결정한다.**

3. 실험환경

3.1 실험 데이터의 구성방법

실험을 수행하기 위해서 먼저 고려해야 할 부분은 두 가지였다. 하나는 단어와 관련되어서 현재 사용 가능한 표준 데이터가 없다는 점이다. 그리고 두 번째 문제로는 연속 필기된 한글 단어의 분할에 관련된 문제이다. 특히 분할의 문제는 매우 어려운 과제중의 하나로서 다른 또 하나의 과제로 생각된다. 따라서 본 논문에서는 포항공대에서 제작한 PE92 한글 필기 데이터를 적절히 활용하였다. 즉, 원하는 단어에 해당하는 낱자를 PE92 데이터에서 추출하여 단어를 구성하여 인식 실험에 사용하였다.

3.2 실험 대상에 대한 조사

이번 실험에서 대상으로 한 단어는 전국 동, 읍, 면을 나타내는 16010개의 단어이다. 그리고 이들 단어들에 쓰이는 낱자의 개수는 모두 354개이다. 따라서 통계적 인식기는 354개 단어 각각에 대한 표준벡터를 가지고 인식작업을 수행한다.

실험에 쓰인 단어의 길이에 따른 분포는 표2와 같다.

단어의 길이	비 율(%)
2	2.3
3	36.7
4	42.8
5	15.3
6	2.9

표 2 단어길이에 대한 분포

위의 표에서 알 수 있듯이 전체 단어중 길이 3, 4에 해당하는 단어가 약 80%를 차지하므로 단어 인식률은 이 두 가지에 의해 결정된다고 볼 수 있다.

4. 실험결과

표3은 총 9000개의 테스트 단어에 대해서 인식 실험을 수행한 결과이다.

후보수 \ 단어 길이	2	3	4	5	6
1	54	63	76	81	97
2	75	75	84	89	97
3	81	80	88	91	97
4	84	83	89	92	97
5	87	85	90	93	97

표 3 실험 결과(단위:%)

위의 실험결과를 분석해 보면 대체로 단어의 길이가 긴 것이 인식률이 높음을 알 수 있다. 이는 단어의 길이가 길수록 연결확률에 의한 제한조건이 많아지면서 낱자에 대한 후보수나 만들어진 단어의 후보수가 대폭적으로 줄어들기 때문에 인식률이 높은 것으로 생각된다. 그리고 단어의 길이가 2인 경우는 끝자가 동, 읍, 면의 세 가지 경우로 끝나기 때문에 실제적으로 첫 번째 낱자에 의해 단어의 인식률이 결정되는 경우가 많기 때문에 인식률이 낮은 것으로 생각된다. 그리고 길이가 3인 경우는 비슷한 단어들이 많이 존재하기 때문에 인식률이 낮은 것으로 생각된다.

후 보 수	인 식 률
1	72
2	82
3	85
4	87
5	89
6	90
7	90
8	91
9	92
10	92

표 4 후보 수에 따른 인식결과(단위%)

표4는 후보수에 따른 인식결과이다. 10후보까지의 인식률은 92%인데 이는 처음 낱자후보의 개수를 80개로 가정했을 때 예상되는 단어의 인식률인 96%에 근접하

는 것을 알 수 있다. 즉 인식 순위에서는 뒤지지만 일정수의 후보 내에 정답이 포함됨을 알 수 있다.

5. 결론

필기 한글 인식의 과제는 다방면으로 연구되어오고 있지만 그 변형의 다양함으로 인해 매우 다루기 어려운 과제로 알려져 있다. 그래서 낱자단위의 인식은 물론 자소 단위의 인식 연구 등이 여러 곳에서 수행되고 있지만 매우 시간이 많이 걸리는 어려운 작업이다.

따라서, 본 논문에서는 인식의 단위를 단어로 하고 단어의 수를 제한하여 인식하는 시스템을 구성하였다. 즉, 인식해야할 낱자와 단어의 수를 제한하고 또 단어 사전에 있는 어휘지식을 활용함으로서 인식률을 극대화 하도록 하였다.

본 논문에 쓰인 낱자 인식기의 인식률은 저조한 편이다. 그러나 후보개념을 적극 활용하고 어휘지식이 학습된 어휘정보 네트워크를 활용함으로서 낮은 낱자 인식률에도 불구하고 높은 단어 인식률을 얻을 수 있었다.

앞으로의 과제는 첫째로 단어인식 네트워크의 개선에 있다. 단어의 10후보까지의 인식률이 92%정도이므로 후보 선발과정에는 큰 이상이 없다고 할 수 있다. 따라서, 단어인식 네트워크의 연결 확률값을 개선하여 후보 순위의 결정 방법을 바꾸는 것이 필요할 것으로 생각된다.

둘째로는 낱자 인식기의 후보 선정과정에 대한 것이다. 현재는 낱자 인식 결과 값에서 상위 80개만 선정하여 단어인식의 자료로 활용하고 있으나 이는 낱자 인식기에서 나오는 인식결과의 일부분을 잃어버리는 결과를 가져온 것으로 생각된다. 따라서 이런 정보를 잃지 않는 방향으로 낱자인식기의 후보선정 방법을 개선해야 할 것으로 생각된다.

앞에서 언급한 것처럼 단어인식 시스템은 주소인식 시스템 개발을 위한 연구의 일부분으로 진행되고 있는 것이다. 현재는 동/읍/면 단위의 단어 16010개에 대해서만 실험을 하였지만 앞으로 시/도 단위 단어와 구/시/군 단위 단어에 대한 각각의 단어 인식 시스템이 구성되고 이 시스템들이 서로 연결되면 각 시스템의 결과가 서로 다른 시스템에 제한의 효과를 주어서 단어 인식 성능이 더 향상될 것으로 기대된다.

참고문헌

[1] 중앙대학교, 기아정보시스템, "필기체 문자인식 기술개발", *정보통신부 제조업 경쟁력 강화사업 연구보고서*, Sep., 1995.

[2] 정재엽, "은닉 마르코프 모델을 이용한 한글 어절 후처리기", *연세대학교 대학원 석사학위 논문*, Dec., 1994.

[3] H. BUNKE, M. ROTH and E. G. SCHUKAT-TALAMAZZINI, "OFF-LINE CURSIVE HANDWRITING RECOGNITION USING HIDDEN MARKOV MODELS", *Pattern Recognition*, Vol.28, No.9, pp1399-1413, 1995.

[4] H.Mou-Yen Chen, Amlan Kundu, and Jian Zhou, "Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network", *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 16, NO. 5, MAY 1994

[5] BUNKE, M. ROTH and E. G. sCHUKAT-TALAMAZZINI, "OFF-LINE CURSIVE HANDWRITING RECOGNITION USING HIDDEN MARKOV MODELS", *Pattern Recognition*, Vol.28, No.9, pp1399-1413, 1995.

[6] LAWRENCE R. RABINER, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Readings in Speech Recognition*, 1988

[7] R. G. Gonzalez, R. E. Woods, "Digital Image Processing", Addison Wesley, 1992