DP 정합을 이용한 필기체 한자 인식

전상엽, 권희용
안양대학교 컴퓨터공학과

Recognition of Handwriting Chinese Characters Based on DP matching

Sang-Yeop Jun, Hee-Yong Kwon
Dept. of Computer Engineering, Anyang University

요 약

온라인 필기체 한자는 동일인의 동일 문자조차도 획수, 획순 및 형태의 변화가 다양할 뿐만 아니라 인식 대상이 방대하여 인식이 매우 어렵다. 또한 한자는 기본 자소의 조합에 의한 글자가 아닌 각각의 글자가 독립적으로 이루어져 있어 연속된 획들 간의 관련도를 파악하기 어렵고 획수도 1획에서 28획까지 다양하게 분포를 한다. 따라서 본 연구에서는 대분류 단계로 시작획 비교를 하고 이어진 세분류 단계에서 문자의 특징으로 방향코드와 특이점을 추출해내고 획수를 고려하여 DP 정합을 하는 2단계 인식 시스템을 제안하였다. 이로써 최적의 속도로 입력한 문자를 찾아낼 수 있도록 하였다.

1. 서론

최근 정보화 사회를 맞이하여 처리하여야할 데이터가 급속히 팽창하고 있어서 자료를 수작업으로 입력하는 것은 많은 시간과 인력의 낭비를 초래하거나 아예 비현실적이기도 하다. 따라서 자료를 자동으로 입력하기 위한 문자 인식에 관한 많은 연구가 국내외적으로 활발하게 진행되고 있다.[1] 특히 최근 급속히 성장하고 있는 휴대형 단말기의 정보 입력 수단으로 키보드를 대신하는 전자펜과 온라인 문자 인식 기술은 그 중요성이 크게 부각되고 있으며, 특히 한자 입력을 위한 온라인 필기 한자 인식은 한정된 자판과 학습의 어려움 등의 이유로 그 필요성이 크게 대두되고 있다. 본 연구에서는 사람이 직접 입력한 필기체 한자를 효율적으로 인식할 수 있는 방법을 제안하고자 한다.

한자는 대부분 직선과 사선으로 복잡한 획의 구조로 조합되어져 있고 한글과 달리 문자의 구조상 획간 교차점이 다양하게 형성되어 있어 특징 추출을 어렵게 한다[2]. 또한 한자 자체의 방대한 문자 집합과 각 문자에 대한 획순과 획수의 다양성으로 인해 필기자 개개인의 다양한 필기 습관에 따라 여러 가지 문자 변형이 존재하게 된다. 이와 같은 변형의 다양성은 온라인 한자 인식 분야에 어려움을 가중시키고 있다. 따라서 이러한 획순, 획수 및 획 모양의 변형을 흡수하여 인식할 수 있는 효율적인 방법이 절실히 요구되고 있다.

기존의 연구로는 원형 정합 방법, 구문 해석 방법, 통계적 방법, 신경망 방법, 퍼지 추론 방법[3], 그리고 특징 분석 방법[4] 등 다각적인 접근 방법으로 연구가 활발히 진행되고 있다. 기존의 온라인 필기체 한자 인식에 관한 연구는 정해진 표준 획순과 획수를 지켜 쓰도록 필기자에게 요구하여 필기된 문자를 인식하는 연구가 주종을 이루었다. 하지만 최근 들어서는 필기 제약을 없애는 방향으로 많은 연구가 진행되고 있다. 여기서 사용하고 있는 동적 패턴 정합은[5] 사람의 필기체와 같이 형태의 변형이 심한 경우에도 유연하게 정합을 한다. 하지만 동적 패턴 정합은 처리에 많은 시간이 소요된다. 더군다나 한자와 그들의 다양한 변형을 포함하는 방대한 문자 집합을 정합해야 하는 문제가 있다.

이와 같은 문제를 해결하기 위하여 최소한의 데이터를 이용하고 또 정합할 때의 조건을 간소화 하여 인식 시간을 크게 줄일 수 있도록 2단계 인식 시스템을 제안하였다. 따라서 최소한의 비교로 최적의 결과를 낼 수 있도록 하였다. 이하 2장에서 동적 정합 방법을 살펴보고 3장에서는 제안된 2단계 인식 시스템

의 구조와 알고리즘을 소개하였다. 끝으로 실험 결과와 결론을 맺는다.

2. 동적 패턴 정합

동적 패턴 정합이란 동적계획법을 이용하여 두 패턴 요소간의 대응을 수행하여 유사도를 계산하는 방법이다. 유사도를 이용하면 복잡하게 변형된 패턴이 원래 어떠한 패턴과 유사한 패턴이었는가를 판정할 수 있다. 동적 패턴 정합은 전체적 및 부분적 특성을 반영하는 특성을 가지고 있어 음성인식 및 영상처리 방법으로도 사용되고 있다. 이 방법은 기준이 되는 패턴과 왜곡되거나 변형된 입력과의 정합 정도를 평가하는 방법이다. 동적 패턴 정합은 수행이 반복적 계산에 의하여 이루어지므로 구성이 간단하며, 최적의 정합 경로 및 최적해를 구할 수 있다. 획변형과 획수가 다양한 한자 에서는 가장 안정적으로 인식할 수 있는 방법으로 DP정합을 선택하였다.

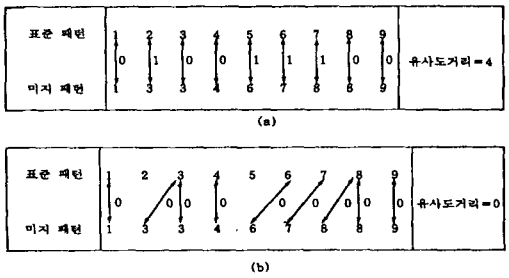


그림1. 선형사상(a)와 신축사상(b)의 예

그림 1에서 사상 w에 의한 패턴 A, B간의 유사도는

$$D_0(A,B) \equiv \min\{D_0(A,B;w) \mid w \in W(I,J)\}$$

이라 정의하고 그 유사도의 합은

$$D_0(A,B;w) \equiv \sum_{i=0}^{I} d(a_i, b_{w(i)})$$

이라 정의한다. 여기서 $D_0(A,B)$

의 계산량은 I번의 덧셈을 한다고 하면

덧셈 : I * |W(I, J)|번

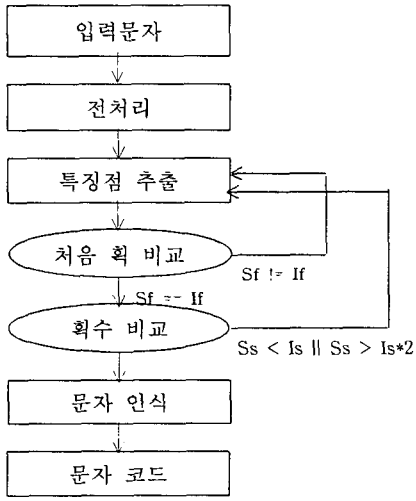
비교 : |W(I, J)|-1번의 계산을 해야 한다. 예를 들어 I=J=25일 경우에 덧셈횟수는 10^{15} 가 된다. 이런 계산의 결과로 프로그램을 수행하는 것은 불가능하다. 따라서 DP정합의 계산을 위한 점화식을 사용한다[6]. 여기서 대칭적인 유사도에 대해서 생각해 봐야한다. 대칭적인 유사도 $D(A,B)=D(B,A)$는 대칭성이 꼭 성립하는 것은 아니라는 것이다. 따라서 이 부분에 대한 계산 과정도 필요하다. 현재 방향코드를 추출하는 부분은 획수에 따라서 방향코드의 크기가 변한다. 따라서 획수가 적은 것과 많은 것을 비교 한다면 대칭성이 맞지 않는다. 그리고 패턴 B의 한 요소 b에 대응하는 패턴 A의 요소가 존재하지 않는 경우와 요소 b가 처음에는 B에 존재하지 않았다가 나중에 삽입되었다는 것을 생각해 볼 수 있다. 이런 경우에 탈락과 삽입을 고려한 DP 정합을 하여야 한다. 프로그램에서는 탈락되었을 때 가중치 2를 주어서 계산을 하였다.

이제 DP 정합을 사용하여 필기체 문자 인식을 하기위한 과정을 살펴보기로 한다. 필기체 한자 인식의 비교 과정은 사용되는 특징의 종류와 밀접하게 연관되어 있다. 본 논문에서는 최소한의 데이터를 가지고 비교를 하게 된다. 이렇게 특징값을 추출하기위해 비교를 하기 전에 전처리 과정을 거치게 된다. 전처리는 필기자의 필기 습관과 잡음 등에 의해 변형과 왜곡을 포함하게 되는데 이런 오류를 최소화하고 안정된 획을 추출하기 위해 적절한 전처리 과정을 거치게 된다. 전처리 과정을 거친 데이터는 방향코드와 특이점으로 특징값을 추출한다. 그리고 그 특징값을 가지고 DP 정합을 하는데 모든 패턴들에 대해서 비교를 하면 시간이 많이 걸린다. 따라서 문자의 획수를 고려하여 첫 획의 일치 여부를 먼저 검사하고 비교 패턴 대상을 걸러낸다. 그러면 비교 횟수가 줄어드는 효과를 낸다.

3. 동적 패턴 정합 제안

본 논문에서 제안하는 필기체 한자 인식 시스템의 전체적인 구성은 그림 2와 같다. 전자 타블렛 펜을 이용하여 입력된 문자 패턴에 대해 잡음과 훅, 중복점 등을 제거하는 전처리를 행한 후 방향 벡터를 이용한 특징점 추출을 한다. 먼저 한자 데이터를 받아오면 문자를 규격에 맞추기 위해서 평활화, 크기 정규화, 공간 정규화를 하였다. 여기서 데이터의 특징을 추출하였다. 특징 추출은 0부터 7까지의 8방향코드를 추출하였고 그 방향코드의 길이를 동적 패턴 정합을 하기 위해서 길이를 정규화 하였다. 이후 특징값을 DP 정합으로 표준 패턴을 찾는다. DP 정합을 할 때 가능한 한 비교 횟수를 줄이기 위해서 처음 획을 먼저 비교한다. 이것은 사람들이 문자를 쓸 때 보통 처음 획의 획순은 잘 변하지 않기 때문에 처음 획을 비교하였고 처음 획이 많이 틀리다면 비교 검색 대상에서 제외하고 다음 문자를 비교하게 된다. 따라서 비교횟수를 줄

일 수 있다. 그 후에 획 수를 고려하게 된다. 획 수는 문자를 쓰는 사람에 따라서 차이가 많이 나는 특징중에 하나이다. 하지만 보통 사람들의 경우는 문자를 쓸 때 획수가 그 문자의 표준 문자 획수의 일정 범위의 배율 한도 내에서 줄어드는 것이 상식이다. 그 이상으로 획수가 줄어들거나 표준 문자보다 획수가 많아지는 일은 드물다. 따라서 표준 문자 집합에서 입력 문자 집합의 획수를 비교 하는데 입력 문자 획수에서 2배 정도 이내의 차이가 나는 표준 문자 집합과 비교를 한다. 따라서 획수 비교를 통하여 DP 정합을 수행하는 횟수를 줄일 수 있다. 그러므로 더 빠른 인식 속도를 보일 수 있다. 따라서 최소한의 비교로 처리 시간을 줄여서 최적의 결과를 낼 수 있도록 하였다.



Sf : 표준패턴에서 처음 획의 특징값
If : 입력패턴에서 처음 획의 특징값
Ss : 표준패턴의 획수
Is : 입력패턴의 획수

그림2 필기체 한자 인식시스템의 전체 구성도

그러면 이제부터 인식을 위한 획 정규화부터 차례대로 알아보자. 획 정규화 과정에는 평활화, 크기 정규화, 공간 정규화 과정이 있다. 먼저 평활화 과정부터 살펴보겠다.

1)평활화

필기자가 문자를 입력할 때 불규칙한 펜 떨림과 획의 시작 부분과 끝 부분에서 매우 짧고 급격한 방향의 전환에 의해 잡음이 발생한다. 이러한 입력의 잡음을 제거하는 정규화 과정을 평활화라 한다. 따라서 문자에서 필요 없는 것들을 제거하게 되었다.

2)크기 정규화

필기자가 쓴 문자는 필기자의 습관에 따라 크기가 다양해진다. 크기가 다양해지면 같은 글자라 하더라도 비교를 할 수 없게 된다. 따라서 필기자가 쓴 문자를 틀에 맞게끔 크기를 맞추어 준다. 이렇게 함으로서 같은 문자들을 같은 크기로 비교할 수 있게 된다.

3)공간 정규화

크기 정규화를 마친 문자는 필기시 필기 속도에 따라서 점들의 위치가 일정하지 않게 되는데 이런 일정하지 않은 점들의 위치를 일정하게 조정하기 위해서 공간 정규화를 해준다. 공간 정규화로 점들간의 등간격을 조정하면 점들간의 위치가 균일해서 특징점을 추출할 때 균일하게 추출할 수 있다.

이렇게 전처리 과정을 거쳐 샘플링된 좌표의 순서쌍을 8방향 코드와 특징점 코드로 변환시킨다. 8방향 코드로 변환시킨 후 문자의 획수와 함께 저장을 한다. DP 정합은 시간이 오래 걸리기 때문에 시간을 단축하기 위해서 비교 횟수를 줄여야 한다. 따라서 여러 가지 제약사항을 둔다.

4) 1단계 정합 과정

첫 번째로 저장된 문자는 먼저 처음 획의 방향코드를 비교해서 일정 범위내에서 같은 방향인지를 검사한다. 같은 방향이 아니라면 비교대상에서 제외시킨다. 예를 들어 입력패턴의 첫 획이 413으로 시작한다면 입력패턴을 크기순으로 정렬하고 중간값 즉 3과 매칭되는 표준패턴을 찾아내서 비교를 한다. 이로서 비교 검색 대상을 30% 이내로 감소 시킬 수 있다. 두 번째 제약사항은 문자의 획수를 비교하는 것이다. 보통 필기자는 문자의 획수를 줄여서 쓰는 것이 보통이다. 따라서 입력패턴보다 획수가 같거나 많은 표준 패턴을 비교 대상에 포함시킨다. 그리고 입력 패턴의 획수는 표준 패턴에서 2배 이상 줄여 쓰는 경우는 드문 경우이므로 입력패턴의 획수에서 2배 이내의 획수를 가진 표준 패턴과 비교를 한다. 예를 들어서 입력패턴의 획수가 5이었다면 표준패턴에서는 5획을 가진 문자부터 비교를 해서 10획까지의 문자를 비교하고 여

기에 포함되지 않는 표준패턴들은 비교대상에서 제외를 시킨다. 그러면 비교횟수를 많이 줄일 수 있다. 이러한 방법으로 DP 정합하는 시간을 단축시킨다. 아래는 이러한 알고리즘을 보여준다.

Algorithm

InData : 입력패턴의 특징값
SData : 표준패턴의 특징값
InData_Stroke : 입력패턴의 획수
SData_Stroke : 표준패턴의 획수
Résultint : 결과값

```
Begin
while(ReadData()){        //입력패턴을 읽어온다.
 if(SData_Stroke< InData_Stroke || SData_Stroke > InData_Stroke*2 || InData[1] != SData[1])
{
        Resultint[i++] = -1;
        continue;
}
if(min > (result=DP(a,s,al,sl)) ){  //DP matching
        min = result
        m_ResultClass = i
}
m_Resultint[i++] = result
End
```

이렇게 첫획과 획수 비교를 통해서 비교 횟수를 줄여 나간다.

5) 2단계 정합 과정

비교횟수를 줄이고 나면 DP 정합으로 방향코드를 비교 해 나간다. 방향코드는 인접 방향코드와 유사도를 계산한다. 하지만 방향코드의 길이가 획수에 따라서 틀려지기 때문에 만약 비교를 할 수 없다면 가중치를 준다. 그리고 방향코드의 길이가 차이나면 차이나는 만큼으로 나누어 줘서 방향코드 길이에 대한 정규화를 하고 비교검색하는 시스템을 만들었다.

4. 결론

한자 인식을 동적 정합(DP)으로 하는 경우에는 한자의 방대한 양과 복잡한 획의 조합, 획수, 획순 등으로 인해 비교하기가 매우 까다롭고 시간이 많이 걸린다. 따라서 간단하고 시간이 적게 걸리게끔 하기 위해 비교 검색 조건에 제약 사항을 두어서 비교 횟수를 줄이는 2단계 인식 방법을 사용하였다. 따라서 기존의 동적 정합보다 빠르게 인식할 수 있게 되었다. 그리고 획수에 제약을 덜 받게끔 설계를 해서 입력패턴의 획수가 줄어들어도 유연성 있게 받아들일 수 있도록 하였다. 하지만 한자의 특성상 비슷한 방향코드와 획을 가지고 있는 것도 많기 때문에 이런 부분에서 인식률이 나빠진다. 따라서 이런 글자들에 대해서 좀더 세부적인 제약 사항을 둔다면 좀더 완벽한 한자 인식기를 만들 수 있다.

앞으로의 연구 과제는 좀더 세부적이고 새로운 알고리즘으로 인식 시간을 줄이는 것이고 현재 1800자로 테스트를 하였지만 더 많은 한자를 전체 표준 집합으로 확장하는 것이다. 이렇게 하려면 특징값을 찾는 방법도 변해야 한다. 비교하기 쉽고 짧은 특징값을 찾아서 비교 횟수를 줄이고 정확한 비교를 통해서 인식률을 높이는 연구가 수반되어야 한다.

[참고문헌]

[1]jeng,B.-S., C.-H. "Chinese character segmentation using the character-gap feature" Applications of digital image processing IX 3834권 pp.262-269

[2] 진원, 김기두 "유닛 재구성 방법을 이용한 PDA영 온라인 필기체 한자 인식" 대한전자공학회 논문지 제 39권 1호 pp.97-107 2002.1

[3] 심동규, 함영국, 박래홍 "DP 매칭과 퍼지이론을 이용한 흘림체 온라인 한글인식" 대한전자공학회 논문지 제 30권 4호 pp.116-129 1993.4

[4]Tseng, YH, Lee, HJ "Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm" Elsevier Pattern Recognition Letters 20권 8호 pp.15

[5]이은주, 박진열, 박재성, 김태균 "구조 정보의 DP 정합에 의한 흘려 쓴 한글의 온라인 인식" 대한전자공학회 논문지 31권 4호 pp.166-174 1994.4

[6] 김상운 "식별 알고리즘을 중심으로 한 패턴인식 입문" 홍릉과학출판사 1997.6