제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

Pocket PC용 한글 매칭 시스템 설계에 관한 연구

이호현*, 조범준*
조선대학교 컴퓨터공학과
mctms@msn.com, bjcho@chosun.ac.kr

A Study on Implementation of Hangul matching System for Pocket PC

Lee Ho-hyun*, Cho Beom-Joon *
*Dept. of Computer Engineering, Chosun University, Gwangju 501-759, Korea

요 약

한글위주의 스크립트를 전자 잉크 데이터(electronic ink data)형태로 Pocket PC에서 사용하기 위한 한글 매칭 알고리즘을 적용한 시스템 설계에 관한 연구이다. 적용된 한글 매칭 알고리즘은 전자 잉크 데이터(electronic ink data)를 스크립트 형태로 변화한 후 이를 모바일 환경의 프로그래밍 기법을 사용하여 시스템을 구현 한다. Pocket PC의 하드웨어적 제약을 고려하여 효율적인 속도를 보장하면서 인식률을 높이기 위해 기본 획을 인식한 후, 획 정보와 획간의 위치관계를 이용하여 자소로 분리된 데이터의 값으로 변환하여 이를 CF메모리상에 있는 통계적 수치 데이터로 저장된 한글 데이터의 값과 비교하여 한글을 인식할 수 있는 시스템 구현에 목적이 있다.

1. 서론

휴대용 컴퓨터에서 할 수 있는 많은 일들 중 기존의 인식 시스템을 굳이 사용할 필요가 없는 몇 가지 이유는 첫째, 사용자들이 Pocket PC 상에서 매일 하는 대부분의 일들은 모두 전자 잉크 영역(electroonic ink domain) 내에서 처리할 수 있는 것들이다. 둘째, 전자 잉크 데이터의 경우 펜으로 입력하는 데이터의 범주를 제한하지 않기 때문에 사용자가 표현할 수 있는 정보의 형태가 다양하다. 셋째, 전자잉크데이터의 값을 이용하여 문자 인식이 필요할 경우 인식을 수행할 수 있는 기능을 구현 수 있다. Pocket PC에서 전자 잉크 데이터의 처리와 인식을 위한 연구는 1990년대 초반부터 이루어져 다양한 연구 결과가 나왔다. 전자 잉크 데이터의 한글 인식을 위한 몇 가지 대략적인 잉크 매칭 알고리즘을 (approximate ink matching algorithms) 이 제안되었으며[3][4] 대용량의 데이터 베이스에서 전자 잉크 데이터를 효율적으로 검색하기 위한 방법으로 은닉 마르코프 네트워크를 이용한 방법과 R-tree를 이용한 방법들이 제안되었다[5][6]. 그런데 제안된 방법들은 영어 스크립트를 주된 입력 형태로 개발되었다. 펜으로 영어와 한글 스크립트를 작성할 때 영어는 연속된 원호의 합으로 분해될 수 있는 글자인 반면 한글의 경우 획이 꺽이는 위치와 획의 방향이 영어에 비해 뚜렷하다는 차이점이 있다. 영어 스크립트 매칭을 위해 제시된 알고리즘을 구현하여[4] 한글 스크립트에 적용한 결과 매칭률이 매우 저조하게 나타났다. 따라서 영어 스크립트에 적용했던 매칭 알고리즘을 한글 스크립트에 그대로 적용하는 것은 무의미하며 한글 스크립트 위주의 전자 잉크 데이터를 사용하기 위해서는 한글의 기하학적인 특성을 고려한 한글 매칭 알고리즘의[7] 적용이 필요하다.

본 연구에서는 한글 매칭 알고리즘을 이용한 한글매칭 시스템 구현에 목적이 있다.

2. 관련연구

2.1 전자 잉크 데이터 개요

전자 잉크 데이터란 펜으로 입력한 문자, 심볼, 그림 등 펜 스트록(pen-sooke) 데이터를 의미한다. 사용자가 펜으로 입력한 데이터를 인식할 필요 없이 잉크데이터 자체로 데이터 값으로 변환 하였다가 처리할 수 있는 시스템을 만들자는 것이다. 만약 인식이 요구되어지면 필요할 때만 인식을 하도록 하는 것이다. 이 방법의 주된 장점은 사용자에게 입력 문자의 종류에 전혀 제한을 주지 않는 다는 것이다. 문자 인식 시스템의 경우 숫자나 알파벳, 한글, 몇 개의 특수 문자만을 사용하도록 하고 있다. 하지만 잉크 데이터 자체를 처리하는 시스템에서는 이러한 제한을 둘 필요가 없다. 물론 한글 전자 잉크 데이터 인식을 위해 패턴 매칭과 연관된 문제들을 해결해야 되지만 기존의 문자 인식에 비해 조금 더 쉬운 문제이다.

2.2 전자 잉크데이터의 구성

전자잉크 데이터는 펜으로 입력한 획들(strokes)의

집합이다. 획은 pen-down에서 pen-up사이에 입력된 점들(points)의 순서 있는 집합으로 정의된다. 잉크 데이터 Ink는 사용자가 펜을 이용하여 입력한 획들 Si의 집합으로 다음과 같이 표현할 수 있다.

$$\text{In}k = < S_j >, 1 \le i \le n$$

$$S_i = < (x_{ij}, y_{ij}, t_{ij}) >, 1 \le i \le m_i \quad \text{(식1)}$$

식 1에서 Si는 사용자가 입력한 i번째 획을 의미하고 원소 Xij .Yij .Tij는 각각 Si의 j 번째 점의 x 좌표 값, y 좌표 값, 그리고 점이 입력되어진 시간 순서를 나타내며, mi,는 Si를 구성하는 점들의 개수를 의미한다 펜으로 데이터를 입력할 때 동일한 사람이 동일한 글자를 입력해도 두 번 이상 쓸 때 똑같이 쓸 수 없기 때문에 잉크 데이터의 정보는 달라지게 된다. 그러므로 잉크 데이터 인식이 정확히 일치하는 한글 데이터 인식은 불가능하다. 따라서 잉크 데이터 인식은 잉크 데이터의 대략적인 매칭(approximate matching)을 통한 인식이라고 표현하는 것이 적절할 것이다. 전자 잉크 데이터의 인식 문제는 두 개의 전자 잉크 데이터 P(pattern), T(text)가 주어져 있을 때 P가 T 내에 포함되어 있는지, 포함되어 있으면 어느 위치에 있는지를 결정해 주는 것이다. 고전적인 스트링 매칭과 다른 점은 P와 T 사이에 완벽한 매칭이 발생하지 않는 다는 것이다. 펜으로 입력한 필기체는 동일한 글자라 하더라도 사용자에 따라 필기하는 행태와 획 순서가 다를 수 있다. 한글 매칭 알고리즘[7]은 이러한 웬 데이터의 특성을 고려하여 필기자 종속(writer-dependant)인식 방법으로 시스템을 개발하였다. 필기자 종속 매칭 알고리즘의 경우 필체가 다르면 동일한 글자에 대해서도 다른 모양의 객체로 간주하기 때문에 필체가 다르면 데이터 베이스내의 동일한 글자를 쿼리로 주더라도 원하는 정보를 찾을 수 없다. Pocket PC와 같은 소형 정보 단말기의 경우 개인전용 단말기라는 특성을 가지고 있으므로 필기자 종속 매칭 알고리즘이 적절하며 이는 필기 형태가 다른 사람들에 대해 보안의 기능도 제공할 수 있다.

휴대용 컴퓨터의 발전이 가속화되고 있지만 아직 데스크 탑 컴퓨터에 비해 하드웨어 및 소프트웨어 환경에 제약이 따른다. 이런 하드웨어 환경을 고려할 때 CF메모리를 이용하면 데이터 베이스가 커지더라도 효율적인 속도를 보장하고 효율적인 한글 인식이 가능하다고 본다.

3. 제안된 한글 매칭 시스템

3.1 한글 매칭 시스템

본 연구에서는 한글 위주의 스크립트를 전자 잉크 데이터 형태로 Pocket Pc에서 사용하기 위한 한글 매칭 알고리즘을 이용하여 시스템을 구현하였다. 한글 매칭 알고리즘을 이용하여 잉크 데이터를 기본 획 단위로 나눈 후 동적 프로그래밍 기법을 적용한다. Pocket PC의 하드웨어적인 제약을 고려하여 효율적인 속도를 보장하면서 인식률을 높일 수 있도록 CF 메모리상에 한글 수치 데이터를 정장하도록 고안되었다. 적용된 시스템은 Pocket PC상의 다양한 전자 잉크 데이터를 이용하여 한글 데이터 값으로 변환하여 주었을 때 원하는 CF 메모리상의 기존 한글 데이터 값을 검색하여 이를 한글로 매칭해 주는 시스템이다.

3. 2 전처리

필기할 때 손의 떨림으로 인한 굴곡, 획의 끝 부분에서 삐침 등이 생길 수 있다. 이런 오류를 평활화, 여파, 삐침제거, 크기 정규화 등의 전처리 과정을 거쳐 보정한다.

$$X_i = \frac{(X_{i-3} + 3X_{i-2} + 6X_{i-1} + 7X_i + 6X_{i+1} + 3X_{i+2} + X_{i+3})}{27} \quad \text{(식 2)}$$

식 (2)과 같이 전후 7점에 대해 가중치를 적용한 평균값으로 현재 점을 대체하여, 획의 굴곡을 완만하게 하는 평활화를 수행하였다. 필기속도에 의한 불규칙한 점들의 간격은 DDA(Dot Density Algorithm)와 같은 여과과정을 통해 일정하게 할 수 있었다. 본 연구에서는 방향벡터를 특징으로 사용했기에 그 외의 전처리 과정은 생략할 수 있었다.

3. 3 특징 추출 과정

특징추출과정은 개선된 획들의 좌표 데이터로부터 자소 인식을 위한 정보로써 특징벡터, 가상벡터[15,16] 등을 추출한다. 입력 획은 필기체의 특성상 정형화되지 않은 길이와 각도를 가지며 여러 가지 형태의 흘림이 가미될 수 있다. 그러므로 입력 획의 좌표 데이터열에서 불필요한 정보의 양을 줄이기 위해 특징점을 인식의 기본단위로 이용한다[12,13]. 그러나 필기 속도나 각도 변화에 따라 입력 획의 특정부분에 여러 개의 특징점이 존재할수 있으므로 고리, 장식선 제거 및 인접한 특징점 제거과정에 의해 주어진 거리 임계치 이내에 군집된 불필요한 특징점들을 제거한다. 특징벡터와 가상벡터는 한글데이터베이스와의 정합을 위한 정보로써 특징벡터는 입력 획에 대한 순서 및 방향정보이며 가상벡터는 자소 내 획간의 위치관계나 자소 간 위치관계에 대한 정보이다[12,13]. 그러나 자소 인식시 후보 자소들이 많이 발생하지 않으면 빠른 처리를 위해 가상벡터는 Tail-Head Vector 만을 이용한다. 또한 자소 분리와 인식을 위해서 획간의 위치정보인 포함관계와 위치관계를 추출한다. 포함관계는 현재 획을 포함하는 최소외접 사각형과 이웃하는 획들의 최소외접 사각형간의 인접한 정도를 나타내는 정보로써 포함, 겹침, 분리로 나누어진다. (그림 1)(a)는 입력문자 '캉'에 대해서 포함관계를 추출하는 예를 보여준다. 점선사각형은 획을 포함하는 최소외접사각형이며 번호는 점선사각형안에 있는 획 번호이다. 즉, 점선사각형 1은 획 'ㄱ'에 대한 전체 크기를 나타낸다. 점선사각형 2는 점선사각형 1에 포함되며 점선사각형 1,3,5는 서로 분리 되며 점선사각형 3과 4는 겹쳐있음을 알 수 있다. 위치관계는 현재 획에 대한 무게중심과 모든 입력 획의 무게중심간의 방향정보로써 자소, 내, 자소 간 위치관계정보로 이용한다. (그림10)(b)는 입력문자 '긱'의 위치관계를 추출하는 예로써 추출된

위치관계정보는 문자를 구성하는 각 자소간의 무게중심들간의 방향정보이다. 이러한 획간의 위치관계는. 입력된 획들을 하나의 곡선으로 생각할 때 획을 구성하는 점들에 대한 곡률을 계산하면 기본 획으로 분리시킬 위치를 찾을 수 있다. 획을 구성하는 임의의 점에 대해 임계치 이상의 곡률을 가지면 그 점에서 획을 분리하였다.

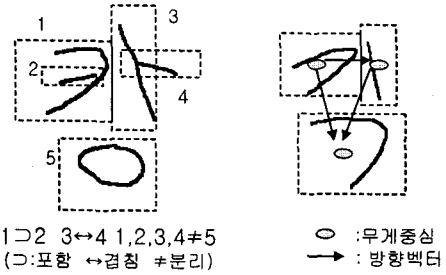

그림 1 획간의 위치 관계 추출

3. 4 기본 획 종류 결정

자소분리는 자소별 인식을 위해 반드시 필요한 과정으로, 본 논문에서는 자소분리와 인식을 병행하여 처리한다. 자소인식은 자소분리에 의해 추출한 각 자소별 획 정보와 한글 데이터베이스에 저장된 자소별 획 정보간의 정합에 의해 자소를 인식하며 인식된 각 자소들의 정보를 전자잉크데이터의 지정된 값으로 변경하여 문자를 생성한다.

자소 분리는 획간의 위치관계를 이용한 순차적 자소 분리와 한글 데이터베이스와의 정합에 의해 각 자소를 구성하는 획 수를 변경함으로써 인식을 수행하는 백트래킹 자소 분리와 한글 데이터베이스와의 정합에 의해 각 자소를 구성하는 획 수를 변경함으로써 인식을 수행하는 백트래킹 자소분리를 이용한다. 순차적 자소 분리는 획간의 포함관계가 분리일 경우 다른 소소의 획이라고 인식하며 겹침이나 포함일 경우 같은 자소 내의 획으로 인식한다. 그러나 자소 내 획간의 포함관계가 분리일 경우 순차적 자소분리는 정확한 자소 분리능력이 미흡하므로 오인식이나 미인식이 발생될 경우 백트래킹 자소 분리를 수행한다. 기존의 백트래킹 방법은 모든 과정을 재처리해야하는 단점을 가진다. 그러나 제안된 백트래킹 자소 분리는 기존에 추출한 획 정보를 기반으로 입력문자가 가지는 전체 획의 개수를 각 자소별 인식 여부에 따라 최적의 획 개수로 분리하여 인식을 수행한다. (그림 2)은 이러한 백트래킹 자소분리에 대한 알고리즘을 나타내며 (그림 3)은 자소 분리와 인식이 병행하여 처리되는 과정을 보여준다. 본 논문은 입력문자나 필체의 형태 정보에 따라 차별적으로 자소 분리를 수행함으로써 빠른 처리속도로 가진다.

Step 1. 순차적 자소분리
Step 2. 초성인식
    if (인식 성공) step 3
    else 백트래킹 적용, 초성 재인식
Step 3. 중성 인식
    if (인식 성공)
        if (종성이 없는 경우)
            자소분리종료, 문자인식
        else step 4
    else 백트래킹 적용, step 3
    if (재인식성공) step 4
    else 백트래킹 적용, step 2
Step 4. 종성인식
    if (인식 성공) 자소분리 종료 문자 인식
    else 백트레킹 적용, step 3
        if(종성재 인식 성공) 자소분리 종료,
            if (종성재 인식 성공) 자소분리 종료,
            문자 인식
            else 백트래킹 적용, step 2
        else 백트래킹 적용, step 2

그림 2 백트래킹 자소분리 알고리즘

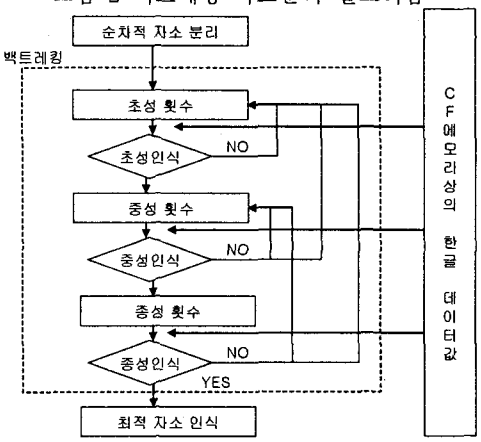

그림 3 자소분리와 인식 처리도

3. 5 유사도 계산

입력된 데이터들을 획의 곡률을 이용하여 기본 획 단위로 분리하는 과정을 거친다. 기본 획은 Pocket PC에서 빠른 매칭 속도를 보장하기 위해 한글의 기하학적인 특성을 고려하여 일곱 개를 지정하였는데 그 [7]과 같은 방법을 사용할 것이다.. 다음 단계로 곡률에 의해 분리된 기본 획들의 종류를 결정하여 획 특징 벡터(stroke feature vector)를 생성한다. 획 특징 벡터는 입력된 데이터를 기본획 단위로 분리한 정보를 가지고 있는 벡터이다. 획 특징 벡터가 생성되면 동적 프로그래밍에 의해 쿼리의 획 특징 벡터와 데이터 베이스 내 잉크 데이터들에 대한 획 특징 벡터들 사이의 거리값을 계산하여 가장 적을 거리값을 가지는 데이터를 CF 메모리상의 데이터 수치값과 비교하여 매칭 결과로 돌려준다.

3. 6 매칭 결과 보고

자소 인식은 CF메모리상의 한글데이터베이스의 수치값의 동기화 과정으로 이루어지며 CF메모리상의 한

글의 구조적 유형정보와 각 자소에 대한 다양한 사용자들의 획 정보 등을 지닌다. 한글의 구조적 유형정보는 중성 인식에서 사용되는 자소 간 결합정보, 모음정보, 획간의 위치관계, 유형에 따른 자소 분류 정보 등을 포함한다. 또한 자소별 인식을 위한 자소 모델은 각 자소에 대해서 일반적으로 사용하는 필체 정보 등을 지닌다. 하지만 심한 곡선이나 자소 간 흘림이 가미된 자소에 대해서는 정확한 정보를 추출하기 힘들기 때문에 직선, 완만한 곡선 정보, 다양한 방향의 필체 유형만을 이용한다. (그림 13)은 한글 데이터 잉크에 대해 계층적으로 인식하는 처리과정을 보여준다.

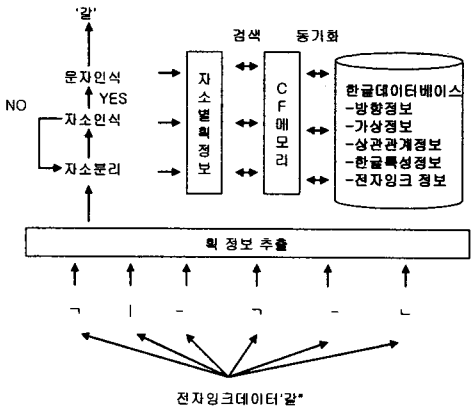

그림 13 계층적 인식 과정

4. 구현 및 고찰

본 시스템은 Windows 2000 환경에서 C++, MFC, WinSock API를 이용하여 구현하여 CF메모리를 제어하였고, CF 메모리상의 한글데이터의 수치 값을 만들기 위해 MS-SQL를 사용하여 일정한 한글 데이터 값을 만들 었다. COMPAQ Pocket PC IPAQ3850를 이용하여 한글 매칭 시스템을 구현하여 테스트 하였다. 현재 본 시스템은 중소기업용 모바일 솔루션의 인증부분에 적용되어 구현되고 있고, 향후 기업내 전자 결제 시스템이나 모바일 전자 결재 시스템에 폭넓게 적용되고 응용될 수 있을 것으로 예상된다.

5. 결론

본 연구에서는 Pocket PC에서 펜으로 입력한 스크립트 데이터를 이용하여 전자 잉크 형태로 저장 및 인식하기 위한 한글 매칭 시스템을 제안하고 구현하였다. 제안된 한글 매칭 시스템은 전처리 과정과 함께 획의 곡률을 계산하여 기본획으로 분리하는 과정을 거친다. 그리고 기본획 종류 결정에 의해 획 특징 벡터를 생성하고 이를 이용한 동적 프로그래밍 기법에 의해 거리값을 계산한다. 거리값을 계산할 때 사용한 편집 연산은 삭제 삽입 교환 연산 외 한글의 특징을 고려한 결합, 분리 연산을 사용하였다. 여기서 만들어진 값을 이용하여 CF 메모리상의 한글 수치 데이터 값을 비교하여 한글을 인식하여 이를 시스템에 적용하였다.

향후 연구 과제는 첫째, CF메모리상의 얼마나 정확한 한글 수치 데이터 값을 가지고 있느냐에 따라 인식의 형태가 바뀌는 것을 알 수 있어 이 문제의 해결을 위해 효율적인 데이터 값을 만들어 시스템의 적용하는 부분 둘째 Pocket PC상의 사용자의 전자 잉크 데이터를 가지고 와서 각각의 사용자의 특성을 파악하여 이를 CF 메모리상의 데이터 값에 적용하는 부분이다.

[참고문헌]

[1] W. Aref, D. Barbara, D. I. PIesti. and A. Tomkins. "Ink as a flrst-class datatype in multlmedia databases," Multimedia Database, Springer-Verlag,1995.

[2] Walid G. AIlaf, Ibrahim Kamel, and Daniel P. IAPIesti, ·'On Handling Electronic Ink," ACM Computing SulVeys, Vol. 27, No 4. pp. 564-567,1995.

[3] IAPIlesti, D. Snd Tomlfins, A" "On the sear-chability of electronic ink," k Froceedings of the International Workshop Front. in Handwriting Recognition, pp. l56-l66, l994.

[4] D.P, Lopresti and A. TOInkins, "Approximate matching of hand-drawn PictogI·am,' In PIneeeding of the Third International Workshop on Frontiers in Handwriting Recognition, pp. lO2-lI1, l993

[5] Walid Aref and Daniel Barbam, "SuppOIting ElectInnic Ink Database," Information Systems, Vol-24, No. 4, pp. 303-326, l999.

[6] Ibrahim Kamel and Daniel Barbara, "Retrieving ElectlOnic Ink by Content," IEEE Proceedings of International Workshop on Multimedja Database Management Systems, pp. 54-61, 1996.

[7] 조 미경, 조 환규, "PDA를 위한 한글 스크립트 매칭 알고리즘," 정보과학회 논문지 제 28권 10호, 2002.

[8] 권요성, 현영빈, "스트링 정합 방법에 기반한 온라인 자소 인식", 한국정보과학회 논문지 제2l권 제5호, pp.750-755, l994.

[9] 싯봉기, 김진형, "은닉 마르코프 모델 네트워크에 의한 온라인 흘림 필기 한글 인식', 한국정보과학회 논문지 제21권 제 9호, pp.1737-1745, 1994.

[lO] Hall g Joon Kim, PyeoLing Kee Kim, "OIrune lecognition of cursive KORan characters using set of extended plimitive sookes and fuzzy functions," Pattern Recognition Letters, Vol-17, pp19-288, 1995.

[12] 이성환, 문자인식이론과실제(I)(II), 홍릉과학 출판사, 1993.

[13] 김대수, 신경망이론과응용(I)(II).하이테크정보, 1993.

[14] 양종원, "DMNN 신경망을 이용한 온라인 한글 필기체 인식에 관한 연구", 전남대학교 석사학위논문, 1997.

[15] 김향미, "순환신경망을 이용한 온라인 문자 인식" 연세대학교 석사학위논문, 1995.

[16] 정기철, 김상균, 이종국, 김상준, "자소 단위의온라인 흘림체 한글 인식", 전자공학회 논문지 제 33권, 제 9호, pp.124--134, 1996. 9.