논문 2010-1-9

필기 입력데이터에 대한 언어식별 시스템의 설계 및 구현

Design and Implementation of a Language Identification System for Handwriting Input Data

임채균*, 김규호**, 이기영***

Chae-Gyun Lim*, Kyu-Ho Kim, Ki-Young Lee*****

요 약 최근, 유비쿼터스 시대로의 도약을 위하여 모바일 기기의 입력 인터페이스에 대한 연구가 활발하게 진행되고 있으며, 기존의 마우스, 키보드뿐만 아니라 필기, 음성, 시각, 터치와 같이 다분야로 세분화되어 새로운 인터페이스가 연구되고 있다. 특히 소형 모바일 기기에서는 크기로 인하여 추가가능한 장치의 제약이 심하기 때문에, 작은 화면에서도 효율적인 입력 인터페이스로 필기 인식이 대두되고 있다. 필기 인식에 대한 이전 연구는 2차원 영상을 기반으로 하거나 벡터로 입력받은 필기데이터를 인식하는 알고리즘이 대부분이며, 필기 인식 알고리즘의 정확성을 향상시키는 연구에만 초점을 두고 있는 실정이다. 또한 실제 필기 입력하는 경우에는 현재 문자가 영문 대/소문자, 한글, 숫자 등의 어느 분류에 속하는지 선택해야하는 문제가 있다. 따라서 본 논문에서는 입력된 필기 데이터를 기반으로 형상분석을 통하여, 영문이나 한글의 여부를 판단하고 언어식별이 가능한 시스템을 제안하였다. 제안 기법은 벡터 단위의 집합으로 필기 데이터를 취급하여 각 벡터 간의 상호관계와 방향성을 분석함으로써 효율적인 언어식별을 가능하도록 하였다.

Abstract Recently, to accelerate the Ubiquitous generation, the input interface of the mobile machinery and tools are actively being researched. In addition with the existing interfaces such as the keyboard and curser (mouse), other subdivisions including the handwriting, voice, vision, and touch are under research for new interfaces. Especially in the case of small-sized mobile machinery and tools, there is a increasing need for an efficient input interface despite the small screens. This is because, additional installment of other devices are strictly limited due to its size. Previous studies on handwriting recognition have generally been based on either two-dimensional images or algorithms which identify handwritten data inserted through vectors. Futhermore, previous studies have only focused on how to enhance the accuracy of the handwriting recognition algorithms. However, a problem arisen is that when an actual handwriting is inserted, the user must select the classification of their characters (e.g Upper or lower case English, Hangul - Korean alphabet, numbers). To solve the given problem, the current study presents a system which distinguishes different languages by analyzing the form/shape of inserted handwritten characters. The proposed technique has treated the handwritten data as sets of vector units. By analyzing the correlation and directivity of each vector units, a more efficient language distinguishing system has been made possible.

Key Words : Handwritten Data, Recognition, Shape Analysis

Ⅰ. 서론

오늘날 전 세계적으로 유비쿼터스 사회를 향하여 전진하고 있으며, 이에 발맞춰 다방면으로 유비쿼터스 관

*준회원, 을지대학교 의료산업학부
**정회원, 을지대학교 의료산업학부 교수
***종신회원, 을지대학교 의료산업학부 교수(교신저자)
접수일자 2010.1.22, 수정일자 2010.2.8

련 기술들이 연구·개발되고 있다. 게다가 장소의 제약이 없이 언제 어디에서나 네트워크로 접속을 가능하게 하기 위하여 다양한 형태의 단말기를 개발하고 있으며, 새로운 입출력 인터페이스도 연구되고 있다.

특히 모바일 기기에서는 작은 크기로 인하여 입력 인터페이스의 추가에 많은 제약이 따른다. 이러한 소형 모바일 기기에서도 효율적으로 적용이 가능한 입력방식으로 펜이나 손을 통한 필기 입력의 분야가 각광받고 있다[1],[2]. 모바일 기기 상의 액정에서 직접 전자펜이나 손으로 입력하는 방식이므로, 입력 장치가 차지하는 부분을 축소하고 화면을 최대화하는 것이 가능하다. 그러나 현재 필기입력 장치는 입력 시에 한글, 영문 대/소문자, 숫자 중에서 어떤 문자를 입력하는지 결정한 후에 입력해야하는 문제가 존재한다.

따라서 본 논문에서는 입력된 필기데이터에서 벡터 간의 상호관계와 방향성을 분석하여 언어식별이 가능한 시스템을 제안하였다. 제안 기법은 각각의 필기데이터를 벡터의 집합으로 취급하여 X, Y 축 방향성과 형태 정보를 이용하여 입력된 언어의 한글 여부를 식별할 수 있도록 하였다. 또한 성능평가를 통하여 제안 기법의 인지 정확도를 나타내었다.

본 논문은 2장에서 필기 입력데이터와 필기 인식에 대하여 설명하고, 3장에서 전체적인 시스템의 설계, 4장에서 시스템 구현을 보이며, 5장에서 성능평가를 하고, 6장에서 결론을 맺는다.

II. 관련 연구

2.1 필기 입력데이터

필기 입력데이터의 종류는 그림 1과 같이 크게 2차원 영상으로 입력되는 경우와 벡터로 입력되는 경우의 두 가지로 분류할 수 있다[3].

(a) 2차원 영상 데이터 (b) 벡터 입력 데이터

그림 4. 필기 입력데이터의 종류
Fig. 4. Types of Handwriting Input Data

입력데이터가 2차원 영상일 경우에는 어떤 방향성이나 순서에 대한 정보를 얻을 수 없어 정적인 이미지에 대한 분석기법이 필요하다.[4] 반면에 벡터 형태로 입력되는 경우에는 (x, y) 좌표계를 통하여 위치, 방향성, 입력순서 등의 정보가 존재하므로 동적인 분석이 가능하다. 따라서 벡터 입력데이터의 경우가 구현이 간단하며 더 높은 정확도를 보인다[5].

다음의 그림 2에서는 벡터 형태로 한글을 입력할 경우, 한글의 자음 및 모음에 대한 필기 벡터의 예를 보이고 있다. 이와 같이 벡터 데이터는 방향이나 위치와 같은 정보를 가지므로 데이터 처리가 용이하다.

(a) 한글 자음의 필기벡터 (b) 한글 모음의 필기벡터

그림 5. 벡터 필기 입력데이터
Fig. 5. Vector Handwriting Input Data

2.2 필기 인식

필기 인식의 과정은 그림 3과 같이 전처리(Preprocessing), 대분류(Pre-classification), 상세분류(Matching), 후처리(Postprocessing)의 네 단계로 구분할 수 있다[6],[7],[8].

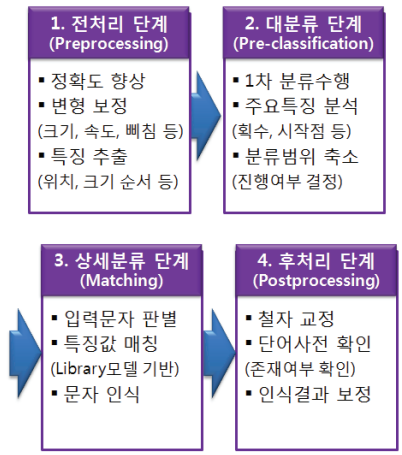


그림 6. 필기 인식
Fig. 6. Handwriting Recognition

전처리 단계에서는 필기에서 크기, 속도, 삐침 등의 변형을 최소화하도록 보정하고, 특징을 추출한다. 다음으로 대분류는 획의 수, 시작점 등의 간단한 특징만을 이용하여 상세분류 단계로의 진행여부를 결정하며, 상세 분류 단계에서는 모델 라이브러리의 특징 값과 입력데이터 특징 값의 거리를 구하고, Linear Matching, Elastic Matching, Neural Network, Hidden Markov Model 등의 알고리즘을 적용하여 사용자가 어떤 글씨를 썼는지 파악한다. 마지막으로 후처리 단계는 인식된 결과를 가지고 사전(Dictionary)에 존재하는 단어인지 확인하여 인식 결과를 수정하게 되는데, 한글의 경우에는 일반적으로 이러한 단계를 거치지 않고 영어인 경우에 주로 사용한다.

III. 시스템 설계

3.1 시스템 흐름도

본 시스템은 그림 4와 같이 벡터 집합의 형태로 입력된 필기데이터에서 벡터를 세분화하며, 각 벡터의 방향성과 위치 정보를 추출하고, 상호 간의 거리를 기반으로 분석하여 언어를 식별하였다.

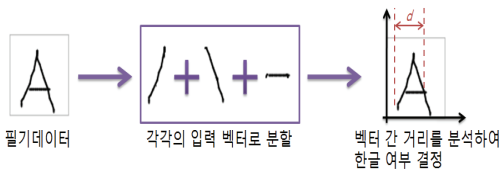


그림 7. 언어식별 처리과정
Fig. 7. Language Identification Process

필기로 'A'를 입력하면 벡터집합은 '/', '\', '–' 의 세 가지 벡터로 구성되게 된다. 각각의 벡터로 분할하여 방향성과 위치 정보를 수집하고, 이를 기반으로 벡터간의 거리를 측정하게 된다. 한글은 여러 벡터로 구성되며 각 벡터 사이의 갭이 존재하므로, 여기에서 측정된 벡터간의 간격이 최소값 이상인 경우 한글로 판정하였다.

그림 5에서는 시스템의 전체적인 흐름도를 나타내고 있다. 먼저 필기 입력데이터를 수신하며, 데이터의 벡터 집합으로부터 각각의 단일 벡터들을 분리한다. 만약 분리된 벡터의 수가 1개라면 한글이 아니라고 인지한다. 기본적으로 한글은 "자음+모음"의 구조를 가지므로 최소한 벡터는 2개 이상이어야 한다. 한편 2개 이상의 벡터를 가진 경우에는 벡터로부터 (x, y)의 좌표와 방향, 입력 순서 등의 특징을 추출한다. 그리고 추출된 특성들을 기반으로 벡터 간의 상호관계를 분석하여 언어의 한글 여부를 식별하게 된다. 기본적으로는 벡터 사이의 유클리디안 거리를 산정하며, 다른 특징 정보와 결합하여 평가한 점수를 임계값과 비교하게 된다. 만약 평가 점수가 임계값보다 크다면 한글로 식별한다.

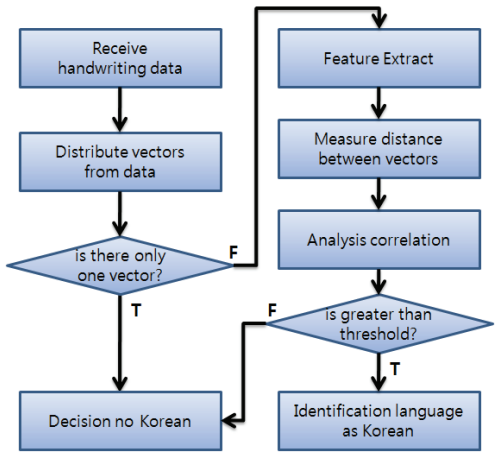


그림 8. 시스템 흐름도
Fig. 8. Syatem Flow Chart

3.2 시스템 구성 모델

입력된 필기데이터에서 수집된 벡터 정보를 기반으로 언어를 식별하기 위해서 사용하는 시스템의 구성 모델은 그림 6과 같이 5-layer의 구조를 보인다.

시스템의 최하위 계층과 최상위 계층은 입/출력을 기반으로 상호작용하는 역할을 수행한다. 최하위 계층인 Data Receiving Layer에서는 필기데이터를 벡터 형태로 입력받아 처리하며, 최상위 계층인 Language Decision Layer에서는 입력데이터로부터 언어의 한글여부를 식별하고 결정한다.

다른 3가지 계층은 실질적으로 필기데이터의 특징 분석을 처리하는 역할을 하는 계층이다. Pre-processing Layer에서는 입력된 필기데이터의 벡터집합에서 필기 입력 시의 오차를 최소화하기 위해 크기, 삐침, 입력속도, 기울어짐 등을 보정하여 일반화한다. 그리고 Vector Extracting Layer에서 각각의 벡터로 분할하고 그 방향성, 위치 정보를 추출한다. 이렇게 생성된 특징 정보를 기

반으로 Distance Analysis Layer에서 벡터 간의 간격을 측정하고, 상호관계를 분석하여 평가 점수를 산정한다. 이 평가 점수가 임계값 이상인 경우에 한글로 인식한다.

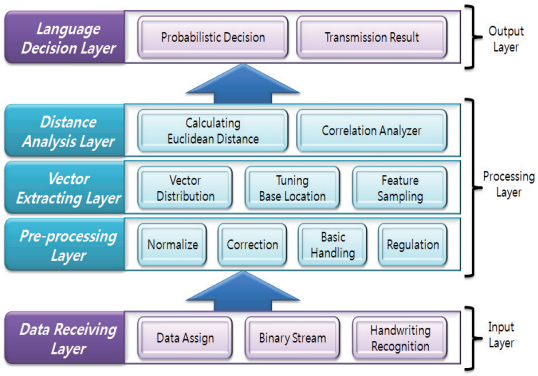


그림 9. 시스템 구성
Fig. 9. System Architecture

Ⅳ. 시스템 구현

본 시스템의 구현은 입력된 필기데이터로부터 필요한 특징 정보를 추출하는 특징추출 알고리즘과 필기의 한글 여부를 판정하는 언어식별 알고리즘의 구현으로 이루어진다. 이 장에서는 각각의 알고리즘의 실질적인 구현에 대하여 기술한다.

4.1 특징추출 알고리즘

필기 입력데이터에서 특징 정보를 추출하는 알고리즘은 그림 7과 같다. 특징추출 알고리즘은 필기 입력데이터를 수신하는 단계로부터 시작한다. 데이터로부터 벡터 집합에 포함된 m개의 단일 벡터로 세분화한다. 그 후에 각각의 벡터로부터 n가지 종류의 특징 (위치, 방향, 크기, 입력 순서, 좌표의 최소값 및 최대값 등)을 저장하게 된다. 여기에서 추출한 특징정보는 각각의 벡터에 저장하여 언어식별에 활용하게 된다.

```
// 필기 입력데이터 수신
RecvData ← Receive all handwriting datas;

// 벡터 세분화
For i =0 to m-1 then
  V[i] ← Distribute vector in current RecvData;
Loop;

// 배열에 각 벡터의 특징 저장
For i =0 to m -1 then
  For j =0 to n -1 then
    V[i].F[j] ← Store vector V[i]'s nth feature
                such as position, direction, size,
                input order, Min and Max for
                x and y, etc;
  Loop;
Loop;
```

그림 10. 특징추출 알고리즘
Fig. 10. Feature Extraction Algorithm

4.2 언어식별 알고리즘

입력된 필기데이터에서 한글여부를 식별하기 위해서 각각의 벡터의 특징정보를 이용한다. 먼저 전체 m개의 벡터에 대해서 각 벡터의 좌표 정보를 이용하여 벡터 간의 유클리디안 거리를 계산하여 기본 정보로 저장한다. 또한 유클리디안 거리를 포함한 n가지의 특징을 기반으로 현재 벡터와 다른 벡터 사이의 상호관계를 분석한다. 이와 같이 특징정보를 분석하여 한글로 식별할 확률을 산정한다.

기본적인 언어식별 알고리즘은 다음의 그림 8과 같다.

```
// 특징 배열에서 벡터간 유클리디안 거리 측정
For i =0 to m-1 then
  E[i] ← Calculate Euclidean distance between
         V[i+1] and V[i];

  // 벡터간의 상호관계 분석
  For j =0 to n-1 then
    PR ← Analysis best correlation for all
         vectors V[i] and all features F[j]
         that covers Euclidean distance E[i];
  Loop;
Loop;

// 임계값 이상이면 한글로 식별
If (PR ≥ threshold) then
  Identify language as Korean;
Else
  Decide no Korean;   // 타 언어로 결정
End;
```

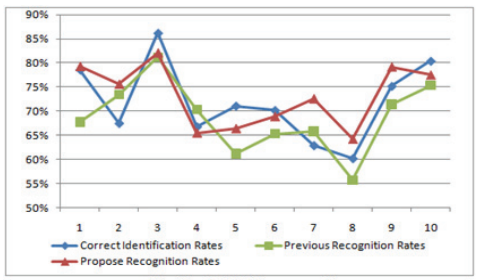
그림 11. 언어식별 알고리즘
Fig. 11. Language Identification Algorith,m

이러한 특징추출 알고리즘과 언어식별 알고리즘으로 구현이 가능한 제안 기법은 필기 인식 과정에서 전처리 단계에 해당하는 역할을 수행한다. 따라서 별도로 입력 문자의 종류를 선택하지 않더라도 한글 여부의 식별을 통하여 기존 필기 인식 알고리즘을 확장하고, 효과적인 성능 향상이 가능하다.

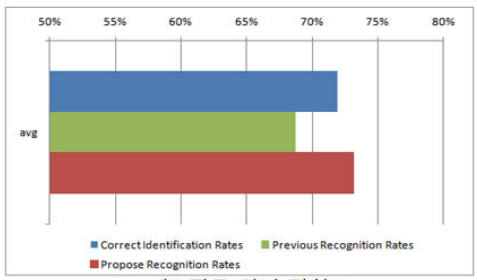
V. 성능 평가

기존의 필기인식 알고리즘은 초기에 어떤 종류의 언어인지 선택하도록 하고 있으며, 알고리즘 수행과정에서는 단순히 선택된 종류의 언어 중에서 가장 유사한 문자로 결정을 내려주는 역할만 담당한다.

그러나 본 논문에서 제안한 알고리즘은 초기에 언어를 선택할 필요가 없이 동적으로 언어를 식별하는 것이 가능하다. 또한 식별하는 과정에서 벡터 간의 간격을 이용하여 한글 여부를 식별하므로, 사용자에 따라 필기의 특징이 다르다고 하더라도 임계값의 수준을 조정하여 처리할 수 있다. 따라서 제안 기법은 적용 범위가 넓고 임계값의 기준을 조정함으로써 정확도를 보정하는 것이 가능하다.



(a) 인지 정확도 그래프



(b) 평균 인지 정확도

그림 12. 인지 정확도 결과
Fig. 12. The Results of Correct Recognition Rates

제안 기법의 성능 평가를 위하여 사용된 시스템에서 하드웨어 사양은 Intel(R) Core (TM)2 Duo CPU T9300 2.50GHz, 4GB RAM이며, 운영체제는 Windows XP SP3를 사용하였다. 10명의 실험자를 대상으로 하여 임의로 선택된 한글 및 영어 문자를 필기로 입력하도록 하였으며, 각 실험자마다 20개씩 문자를 입력하도록 하였다. 필기 입력은 실험자의 습관이 반영되므로 소수의 입력데이터를 사용하더라도 충분히 실험자의 특성이 반영 가능하다. 총 10회의 실험을 수행하였으며, 각 실험마다 200개의 문자 (10명×20개 문자)를 이용하였다.

그림 9는 전체 10회의 실험 결과를 기반으로 제안 기법의 인지 정확도를 나타낸 그래프이다. (a)는 10회의 실험에서 제안 기법이 한글여부를 올바르게 식별한 비율과 기존 기법 및 제안 기법의 인지 정확도를 보인다. 또한 10회 실험의 평균적인 비율을 (b)에 나타내었다.

그림 9의 실험 결과를 통하여 제안 기법이 올바르게 한글여부를 식별하는 비율은 최소 60%~최대 86%로 평균 72%이다. 또한 기존 기법의 인지 정확도가 최소 61%~최대 81%로 평균 69%인데 비하여, 제안 기법의 인지 정확도는 최소 64%~최대 82%로 평균 73%로써 약 6.4% 정도 향상된 결과를 보였다.

실험에서 인지 정확도는 약 73%의 수준으로 현재까지는 높은 수준의 결과를 나타내고 있지 않는다. 이는 알고리즘에 사용된 임계값이 실험적인 요소로부터 결정된 것이 원인이라고 추정된다. 그러나 더 많은 종류의 필기 습관을 학습하여 인지 알고리즘을 보강함으로써 정확도의 향상이 가능할 것으로 판단한다.

VI. 결론

최근에 다양한 단말기에서 필기입력 기법이 유용하게 활용되고 있으며, 여러 필기 인식 알고리즘이 개발되어 있다. 그러나 필기입력 시 한글, 영문 대/소문자, 숫자 등의 언어를 미리 선택하고 입력해야만 하고, 자동적으로 언어를 식별하지 못하는 문제가 발생한다.

따라서 본 논문에서는 벡터 형태로 입력받은 필기데이터를 분석하여 언어식별이 가능한 시스템을 제안하였고, 각각의 벡터 사이의 간격을 분석하여, 임계값 기준 이상일 경우에 한글로 식별할 수 있도록 하였다.

향후에는 벡터 사이의 간격을 판별하는 임계값을 동

적으로 적용하여 알고리즘의 정확도 향상을 연구하고, 더 다양한 언어를 식별할 수 있도록 확장시킬 것이다.

참고문헌

[1] 이성훈, 김진형, "베이지안 망을 이용한 필기 문자 인식 및 생성", 한국정보과학회, *정보과학회지*, 제24권 제12호, 56-62쪽, 2006년.

[2] 조미경, 조환규, "PDA상에서의 한글 필기체 매칭 알고리즘", 한국정보과학회, *정보과학회논문지: 소프트웨어 및 응용*, 제29권, 제9·10호, 684-693쪽, 2002년.

[3] Jia Zeng, Zhi-Qiang Liu, "Type-2 Fuzzy Markov Random Fields and Their Application to Handwritten Chinese Character Recognition", *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 3, pp. 747-760, June 2008.

[4] Huaigu Cao, Venu Govindaraju, "Handwritten Carbon Form Preprocessing Based on Markov Random Field", In *CVPR07*, pp. 1-7, 2007.

[5] Kai Ding, Zhibin Liu, Lianwen Jin, Xinghua Zhu, "A Comparative Study of Gabor Feature and Gradient Feature for Handwritten Chinese Character Recognition", *Proc. of ICWAPR*, pp. 1182-1186, November 2007.

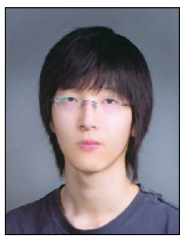
[6] http://www.dt.co.kr/contents.html?article_no=2007072302011832718002

[7] Cheng-Lin Liu, "Normalization-Cooperated Gradient Feature Extraction for Handwritten Character Recognition", *IEEE Trans. PAMI*, vol. 29, no. 8, pp. 1465-1469, August 2007.

[8] Marcus Liwicki, Horst Bunke, "Feature Selection for On-Line Handwriting Recognition of Whiteboard Notes", *Proc. of the Conf. of the Graphonomics Society*, pp. 101-105, 2007.

저자 소개

임 채 균(준회원)



• 2007년~현재 을지대학교 의료산업학부 의료전산학전공 학생

<주관심분야 : u-Healthcare, 유비쿼터스, GIS, 영상처리 등>

김 규 호(정회원)

• 제 9 권 3호 참조

• 1992년~현재 : 을지대학교 의료산업학부 부교수

• 2007년~현재 을지대학교 RIC(지역혁신센터) 부소장

<주관심분야 : u-Healthcare, 유비쿼터스, USN 등>

이 기 영(종신회원) : 교신저자

• 제 9 권 3호 참조

• 2009년~현재 한국인터넷방송통신TV학회 협동이사

• 1991년~현재 을지대학교 의료산업학부 부교수

<주관심분야 : u-Healthcare, 공간 데이타베이스, GIS, LBS, USN, 텔레매틱스 등>