혼용문서에서의 유사문자 분류

문 경애, 지 수영, 오 원근
한국과학기술연구원 시스템공학연구소

The Similar Character Classification in the Mixed Document

Kyung-Ae Moon, Su-Young Chi, Weon-Geun Oh
Systems Engineering Research Institute / KIST

요 약

본 논문에서는 혼용문서에서 문자들의 유사성으로 인해 발생하는 오인식문자를 줄이기위해 대분류 단계에서 유사문자군을 찾고 이들 사이의 유사도를 계산, 분류하는 유사문자분류 방법을 제안 하였다.이 방법은 유사문자군내의 각 문자마다 그 문자만이 갖는 고유한 요인과 그 문자를 제외한 나머지 문자일 가능성이 있는 요인을 찾아 입력문자와 비교하여 유사도가 가장 큰 문자를 인식문자로 선택하는 알고리즘이다.

또한, 인식후 오인식된 문자들에 대해 특징사전의 갱신을 통하여 인식률을 향상시켰다.

I. 서 론

문자인식은, Computer를 이용한 총합적인 기술로서 궁극적인 목적을 문서입력의 자동화에 두고있다. 이러한 문자인식의 핵심기술은 인식방식의 개발에 있음은 두말할 여지가 없으며, 지금까지 국.내외에서 인쇄체문서인식에 관한 연구가 활발히 진행되어 많은 논문이 발표되었고, 또한 상용화된 문자인식 시스템도 나오고 있다[1]. 현재까지 진행된 인쇄체 한글 문서인식에 관한 대부분의 연구는 단일폰트, 단일크기 인쇄체인식이고 최근들어 다중폰트, 다중크기 인쇄체 인식에 대한 연구가 활발히 진행되고 있다[2]. 그러나, 우리가 취급하고 있는 문서는 한글뿐아니라 영문,한자등 다양한 문자들이 다양한 크기와 글자꼴로 혼용되어 있기때문에, 이러한 형태의 문서에서 사용되고 있는 글자꼴을 알 수 없을때 인식은 그리 쉽지 않다. 또한, 각종 글자꼴의 통계적인 특징들이 조금씩 다르기 때문에, 글자꼴의 종류에 무관한 모든경우에 적용할 수 있는 공통적이고 구체적인 특징을 발견하기란 쉽지않다. 혼용문서 인식시스템개발의 문제점으로는 효율적으로 이 문서에 적용시킬수 있는 강력한 알고리즘을 구현하기가 어렵다는 것이다.

본 논문에서는, 혼용문서내의 유사문자분류를 위하여 각 특징들을 집적하여 유사문자인식을 하는, 특징집적에 의한 유사문자 분류 알고리즘을 제시한다. 이 방법은, 대분류를 통하여 유사문자군을 생성하고, 그 유사문자군내의 각각의 문자마다 갖는 그 문자만의 고유요인과 그 문자를 제외한 다른 문자일 가능성이 있는 요인이면서 그 문자에는 없어야 할 요인을 추출하고, 이 두요인을 이용한 입력패턴의 후보문자들과의 유사도를 계산하여 가장 큰 유사도 값을 갖는 후보문자를 인식문자로 선택하는 알고리즘이다.

II. 특징집적에 의한 유사문자인식

본 논문에서는, 문자분류를 위하여 다단계분류 방법을 적용하여 분류의 효율성을 높였고, 이단계에서 생성된 후보문자군들의 유사성을 계산하여 인식 문자를 선택하는 유사문자분류 알고리즘을 소개한다. 먼저 본인식시스템의 전체 흐름도는 그림 1과 같다.

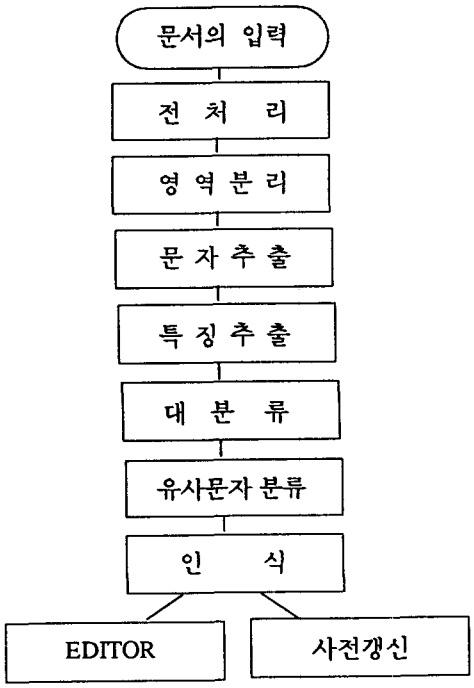


그림 1. 문서 인식 시스템의 흐름

2.1 대분류

본 장에서는 본 문서인식 시스템의 대분류에 대하여 간단히 소개한다. 본 연구에서는, 단순한 문자특징을 이용 짧은시간에 대상문자를 유사문자군으로 일차 분류해 냄으로서 전체 시스템의 처리시간 단축 및 정확도의 향상을 가져오는 다단계 분류방법을 사용했다.

1차 분류로는 네변국소영역에 대해 문자선량에 비례하는 code를 도입하는 4변-code법을 이용하고, 2차 분류로는 문자패턴을 주변으로 부터 본 대체적인 윤곽형상의 특징을 찾는 주변거리분포를 이용한 방법이며, 3차 분류로는 문자의 전체적인 특성을 반영하는 mesh특성을 사용하여 분류하는 pattern 조합법을 이용했다. 분류단계를 거치면서 최종으로 얻어진 후보문자들을 입력패턴에 대한 유사문자군으로 선택한다.

2.2 유사문자 분류

Pattern 조합법과 구조해석법으로 크게 구별되는 인식방식중에서도, pattern 조합법은, pattern의 전체적인 관측에 의해 입력 pattern과 기준 pattern을 조합, 양자의 유사성을 평가하는 방법으로서 pattern의 관측방법이나 유사성의 척도에 무엇을 이용하는가에 따라 여러가지 방법이 제안되고있다. 이 방법은 주로, 인쇄체 문자인식에 적용되고 있으나 일부는 예외적으로 필기체 문자인식에도 적용이 되어, 좋은 결과를 얻고있다[3],[4].

Pattern 조합법은 일반적으로, 문자의 부분적인 변형에 강하고 인식 대상문자의 확장이 용이한점, 사전개발의 자동화가 가능하다는 등의 장점이 있으나, 문자의 전체적인 변형에 대처하기 힘들며, 특히 유사문자의 판별능력이 매우 약하다는 단점이 지적되고있다. 이에, 최근에는 유사문자를 분류하는 방법으로 구조해석법과의 결합 혹은 양방식의 중간적인 방법을 도입하는 경우도 있다[5],[6].

유사문자의 형태는, 인식방식에 따라 약간씩 다르며, 사람이 보아도 구별이 힘든 문자는 어떠한 인식방법을 써도 결과가 좋지않은것이 일반적이다.
본 연구에서 대분류 단계를 거쳐 얻어진 유사문자군의 예를 그림 2에 나타내었다.

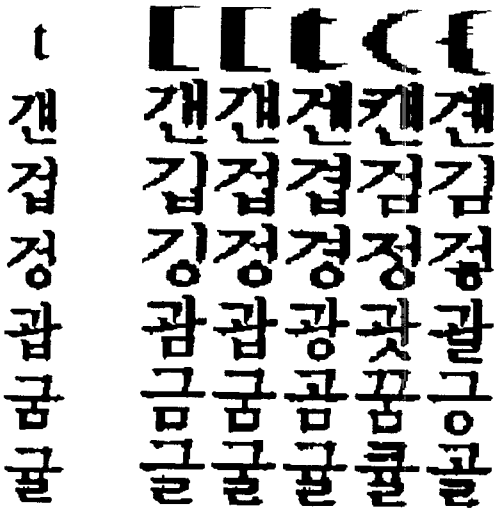


그림 2. 유사문자군

유사문자를 구별하는 방법의 근본적인 원리는, 유사문자간의 상이한 pattern을 사전으로 등록시켜, 인식단계에서 유사문자를 구별하는것이 일반적이다[7],[8]. 그러나, 이러한 방법은 인식방법에 따라 혹은 인식 대상문자의 종류 및 상태에 따라 다르게 얻어질수있는 상이 pattern의 변화에 약하다는 단점을 지니고있으며, 유사문자군의 형태도 바뀔수있어 혼용문서에의 적용은 어렵다.

본 논문에서는, 이러한 점을 고려하여, 유사문자간의 유사성 및 상이성을 특징으로 집적하여, 유사문자를 구별하는 새로운 방법을 제안하였다.

문자 A의 각종 변형(size,폭,shift,회전,잡음 등)을 포함하고있는 문자군을 A = {A}, A내 문자간의 상이함을 XOR(A), 문자군 전체의 포함부분을 OR(A), A내 문자간의 공통부분을 AND(A)라 하면,

$$XOR(A) \supseteq \{A - AND(A)\}$$

$$OR(A) = AND(A) \cup XOR(A) \quad \text{.....(1)}$$

이 성립한다. 즉, OR(A)는 문자 A일 최대범위를, AND(A)는 문자 A일 최소범위를 나타낸다. 모든 인식 대상문자에 대해서 이러한 특징을 계산, 사전으로 등록한다.

다음은, 대분류 과정을 거쳐 얻어진 유사문자군(5자)간에 자기자신일수밖에 없는 (+)요인을,

$$A_+ = AND(A) - OR\{\overline{A}\} \quad \text{.....(2)}$$

자기자신이어서는 안될 (-)요인을,

$$A_{-} = AND\{\overline{A}\} - OR(A) \quad \text{.....(3)}$$

로 하여 각 문자에 대해 계산한다.

인식단계에서는, 미지의 입력문자와 유사문자군내의 각문자의 특징(A_{+}, A_{-})을 각각 비교, 가장 높은 상관관계를 갖는 문자가 인식문자로 선택된다. 이때 학습된 문자 data에 대해서는, +요인과 -요인쪽에서 동시에 높은 상관관계를 나타내어 타 문자와 확실하게 구별되며, 학습이 안된 data에 대해서도 새로운 특징값을 사전에 재등록, 인식율을 향상시킬수있다.

본 방법은, 유사문자간의 상이한 pattern을 미리 등록 할 필요가 없이 인식과정에서 얻을수있기 때문에, 대상문자의 종류나 상태에 제한을 받지않는다. 또, pattern 조합법의 가장 큰 장해요소인 문자의 변형를 흡수, 혼용문서에서의 유사문자를 구별하는 특징으로서 유용하게 사용할수 있다.

2.3 사전의 갱신

본 인식시스템에서는 문자인식후 오인식된 문자에 대해서 특징벡터들과 유사문자인식시 필요한 특징을 재학습시킴으로써 기존의 사전을 갱신시키고, 새로이 갱신된 사전으로 인식을 수행하므로써 인식률을 향상시킨다.

사전의 갱신과정은 대분류에서 참조될 기존의 특징벡터값에 가중치 W를 주고 입력문자에 대해 생성된 특징벡터값을 더해 W+1로 나눔으로서 새로운 평균특징벡터값으로 갱신하고, 유사문자인식에 필요한 특징들 즉, 기존의 and_image와 or_image는 재 학습시킬 문자영상과 각각 and연산, or 연산을 수행하여 새로운 and_imge와 or_image로 갱신한다.

III. 실험 및 결과분석

본 논문의 인식 알고리즘은 PC/486과 Microsoft C를 이용하여 MS Windows상에서 구현하였으며 300dpi 의 resolution으로 scan한 이진 문자영상을 입력문자영상으로 하였다. 인식대상 문자는 한글, 한자, 영문 및 특수문자를 대상으로 하였으며, 다양한 문자크기의 20set을 학습하여 대분류 단계에서 참조될 특징벡터들을 구하고 유사문자인식단계에서 참조될 20set의 and_image와 or_image를 다음과 같이 구한다.

$$or\_image(i,j) = \sum_{k=1}^{n} S_k(i,j)$$

$$and\_image(i,j) = \prod_{k=1}^{n} S_k(i,j) \quad \text{.....(4)}$$

여기서 i,j는 정규화 문자영상의 x,y축 size이고, $S_k(,)$ 는 k번째 학습된 문자set이고, n은 총학습된 set수 이다.

대분류 단계에서 계층적 분류를 하므로서 처리시간을 단축하고, 여기서 얻어진 유사문자군을 가지고 본 논문에서 제안한 유사문자분류방법을 통해 인식률을 높인다.

그림 3은 입력문자패턴 '농'에 대한 대분류후 얻어진 유사문자군'능','농','뇽','눙','늉'을 보여준다. 대분류후 얻어진 유사문자군의 각 문자에 대한 거리 계산값은 '능'은 130, '농'은 132, '뇽'은 199, '늉'은 207로 '능'의 값이 가장 작으므로 유사문자분류 알고리즘 수행전의 인식결과는 오인식을 발생한다. 그러나 유사문자 분류 과정을 통해 유사문자군중에서 입력패턴에 더욱 가까운 문자를 선택할 수 있게 하므로써 오인식을 줄일 수 있다.

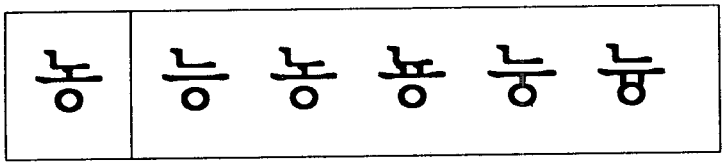


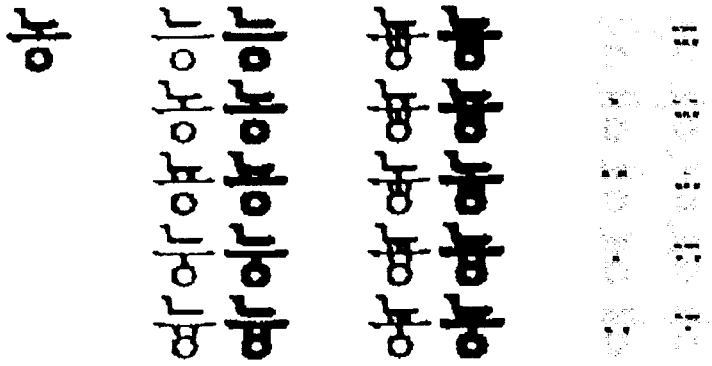
그림 3. 입력패턴 '농'에 대한 유사문자군

그림 4는 유사문자군내의 각문자 마다의 +요인과 -요인을 계산하는 과정을 나타내는데 , (a) 는 입력문자의 패턴이고, (b)는 각 문자의 and_image이고, (c)는 각 문자의 or_image이다. 또한 (d)는 입력패턴에 대한 유사문자군중 자신을 제외한 나머지 문자들의 and_image를 or 연산을 수행한 결과이고, (e)는 입력패턴에 대한 유사문자군중 자신을 제외한 나머지문자들의 or_image를 or 연산한 결과이다. (f)는 입력문자와 +요인과의 and 연산을 수행한 결과이고 (g)는 입력문자와 -요인과의 and 연산을 수행한 결과이다.

(f)와 (g)의 수행결과로 다음과 같이 유사도(δ)를 계산하여 maximum δ_k를 갖는 문자를 인식문자로 한다.

$$\text{유사도}(\delta_k) = \sum_i \sum_j C(i,j)\{P_k(i,j) - M_k(i,j)\} \quad \text{.....(5)}$$

여기서 C(,)는 이진 입력문자패턴이고, P_k(,)와 M_k(,)는 유사문자군에서 k번째 문자에 대한 +요인과 -요인이다.



(a) (b) (c) (d) (e) (f) (g)

그림 4. 유사문자군의 각 문자의 유사도 계산 과정

그림 2와 같이 수행한 결과 유사문자군에서의 각 문자의 유사도는 '능'이 -17, '농'이 18, '뇽'이 -18, '눙'이 -42, '늉'이 -46으로 '농'의 유사도가 가장 큰것으로 나타나 유사문자분류 알고리즘 수행전의 오인식을 보완하여 인식률을 높일 수 있었다.

표 1은 유사문자분류 알고리즘 수행 전,후의 인식률을 보인다.

	수행전	수행후
학습하지 않은 문자 set	95.8 %	98.32 %
학습된 문자 set	96.2 %	98.94 %

표 1. 유사문자분류 알고리즘 수행 전.후의 인식률 비교

실험에 의해, 학습하지 않은 문자 set의 경우 대분류만을 통한 인식결과는 95.8%의 인식률을 얻었고, 오인식 문자에 대해 학습과 유사문자분류 알고리즘을 수행하므로써 98.94%의 인식률을 얻었다.

IV. 결 론

본 논문에서는, 혼용문서내의 유사문자분류를 위하여 각 특징들을 집적하여 유사문자인식을 하는, 특징집적에 의한 유사문자 알고리즘을 제시하였다.

본 방법은, 유사문자간의 상이한 pattern을 미리 등록 할 필요가 없이 인식과정에서 얻을수있기 때문에, 대상문자의 종류나 상태에 제한을 받지않는다. 또, pattern 조합법의 가장 큰 장해요소인 문자의 변형를 흡수, 혼용문서에서의 유사문자를 구별하는 특징으로서 유용하게 사용, 인식율을 향상시킬수 있다.

그러나, 처리시간 및 system의 효율성을 고려할때, 4단계의 인식과정을 세밀한 분석과정을 통해 단축시킬 필요가 있으며, 또, 보다 범용성을 지닌 문자인식 시스템으로 발전시키기 위해서는, 다양한 종류의 문자(잡음, 기울기 등을 포함한 문자)에도 본 방법을동시에 적용할수있도록 연구되어야 할것으로 사료된다.

참고문헌

[1] READEX D 1.0, (주) 코테크 (1993)

[2] 권재욱,조성배,김진형," 신경망 기법을 이용한 다중크기 및 다중활자체 한글문서의 인식 ", 제3회 영상처리 및 이해에 관한 워크숍 논문집,pp.129-136,1992.2

[3] 飯島 泰藏," Pattern 認識 ", Corona社 日本 (1973)

[4] 森 健一 監修," Pattern 認識 ", 日本 電子情報通信協會, Ohm社 (1988)

[5] 安田, 藤澤," 文字認識을 위한 相關法의 改良 ", 日本 信學論, Vol. J62-D, No.3, pp.217 - 224(1979)

[6] 入江, 矢口," 2 * 2 近傍特徵에 의한 筆記體 漢字認識 ", 1988 日本 信學春季全國大會, D-453

[7] N. Sun, T. Tabara, H. Aso and M. Kimura," Printed character recognition using directional element feature ", IEICE of Japan, Vol. J74-D-II, No.3, pp.330-339(1991.3)

[8] S. M. Kang, S. W. Hwang, Y. M. Yang and D. J. Kim," Korean chacter recognition using directional information of character contour ", proc. of 2nd Pacific Rim International Conference on Artificial Intelligence, pp.1198-1202, 1992 Seoul