필기체 수표 금액 문장에서의 단어 분리를 위한 공간적 간격 추정

김인철* · 김경민**

Spatial Gap Estimation for Word Separation in Handwritten Legal Amounts on BAnk Check

In-cheol Kim* · Kyoung-min Kim**

요 약

본 논문에서는 수표 영상 내의 필기체 문장 금액을 개별 단어로 분리하기 위해 연결 성분 간의 공간적 간격을 효과적으로 측정하는 방법을 제안하였다. 기존의 간격 측정 방법들은 과대추정 또는 과소추정과 같은 문제를 내포하고 있어 무제약적 필기 스타일에 의한 문자의 왜곡과 필기 공간의 제약으로 인한 연결 성분 간 중첩 및 접촉, 그리고 단어 간 또는 문자 간 간격의 불규칙성이 심하게 나타나는 문장 금액에 적용하는데 많은 어려움을 가진다. 본 논문에서는 기존의 측정 방식들을 수정 보완하여 측정 과정에서의 오류를 줄임으로써 단어 분리 성능을 향상시키고자 하였다. 단어 분리 실험 결과로부터 수정된 각 거리 측정법이 대응되는 기존의 측정법에 비해 2-3% 정도 향상된 단어 분리 성능을 보임을 확인하였다.

ABSTRACT

An efficient method of estimating the spatial gaps between the connected components has been proposed to separatethe individual words from a handwritten legal amount on bank check. Owing to the inherent problem of underestimation or overestimation, the previous gap measures have much difficulty in being applied to the legal amounts that usually include the great shape variability by writer's unconstrained writing style and touching or irregular gaps between words by space limitation. In order to alleviate such burden and improve word separation performance, we have developed a modified version of each distance measure. Through a series of word separation experiments, we found that the modified distance measures show a better performance with over 2~3% of the word separation rate than their corresponding original distance measures.

키워드

단어 추출, 연결 성분 간 간격 추정, 군집화에 의한 간격 분류, 필기체 문장금액

Ⅰ. 서 론

필기체 문자열 또는 문장에 대한 인식 연구는 필기자의 무제약적 필기 습관에 따른 문자의 심한 왜곡 외에도 문장 내에서의 단어(word) 및 문자(character) 분리 그리고 인식 어휘 수 문제 등으로 인해 그 수행에 많은 어려움을 가지고 있다. 따라서 현재까지 대부분의 필기체 인식 연구는 고립 단어 및 숫자에 대한 인식 문제에 주안을 두고 있다[1]. 그러나, 우편 봉투상의 주소 인식, 수표 상의 문장 금액(legal amount) 인식 등과 같이 제한된 어휘를 가지는 일부 응용 분야를 중심으로 필기체 문자열에 대한 인식 연구가 지속적으로

* CENPARMI, Concordia University, Montreal, Canada
** 교신저자 여수대학교 공과대학 전기및반도체공학과

접수일자 : 2005. 3. 30

진행되고 있다[2][3].

필기체 문자열 인식을 위한 가장 일반적인 접근 방식은 주어진 문자열 영상으로부터 개별적인 단어를 먼저 분리하고 이를 인식한 후에 문맥 정보 등을 이용하여 전체 문자열을 해석하는 것이다. 문자열 영상에서의 단어 추출과 관련된 기존의 연구에서는 특정한 공간적 측정법에 기반하여 인접한 연결 성분(connected component) 간의 간격(gap)을 계산하고 이들을 문자 간 간격(inter- character gap)과 단어 간 간격(inter-word gap)으로 분류하는 기법이 많이 적용되어왔다. Seni [4] 등은 필기체 문장에서 단어를 추출하기 위해 여덟 종류의 거리 측정법을 제안하였다. 이들 중에서, 각 연결 성분을 둘러싸는 최소 사각형(bounding box) 간의 수평 거리를 계산하는 BB 방법과 연결 성분 간의 최소 유크리디언 거리(minimum Euclidean distancce) 및 최소 런 길이(minimum run-length)를 사용한 RLEH 방식이 가장 좋은 성능을 보여준다. Mahadevan [5] 등은 각 연결 성분을 둘러싸는 최소 다각형(convex hull)을 이용하여 서로 인접한 연결 성분 간의 간격을 추정하는 CH 방법을 제안하였다. 전술한 세가지 거리 측정법, BB, RLEH, CH는 모두 간단하면서도 효율적으로 연결 성분 간 간격을 측정할 수 있다는 장점을 가지지만 과대추정(over estimation) 또는 과소추정(underestimation) 과 같은 근본적인 문제를 내포하고 있어 문자 분리 과정에서 많은 오류를 초래할 수 있다.

본 논문에서는 기존 거리 측정법들의 특성을 자세히 분석하고 과대추정 또는 과소추정과 같은 측정 오류를 줄이기 위해 각 측정법들을 수정 보완하였다. BB 방법의 경우에는 수평적으로 돌출된 머리 또는 꼬리 부분에 의해 주로 발생되는 과소추정 문제를 줄이기 위해 연결 성분을 둘러싸는 최소 사각형의 좌우 경계선을 조정하였으며, RLEH 방법에 대해서도 이러한 좌우 경계선 조정 개념을 적용하여 연결 성분의 윤곽선으로부터 직접 거리를 측정함으로써 발생되는 측정의 민감도를 둔화 시키고자 하였다. CH 방법에서는 인접한 두 연결 성분 간의 거리를 수직 이등분 된 좌측 연결 성분의 우측 영역과 우측 연결 성분의 좌측 영역을 각각 둘러싸는 최소 다각형 사이의 거리로 정의함으로써 기존의 CH 방법이 가지는 과대추정 문제를 최소화하고자 하였다.

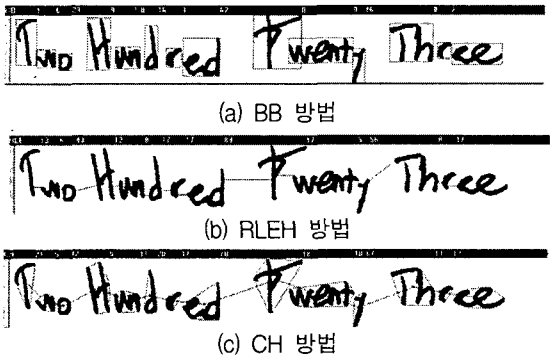
실험에서는 일반적인 필기체 문자열이 가지는 필기자의 무제약적 필기 스타일에 의한 필기 문자의 왜곡 외에도 필기 공간의 제약으로 인해 연결 성분 간 수평적 중첩(overlapping) 및 접촉(touching), 그리고 단어 또는 문자 간 간격의 불규칙성이 심하게 나타나는 수표 내 필기체 문장 금액에 대한 단어 분리 문제에 수정된 거리 측정법을 적용하고 그 성능을 기존의 측정 방식과 비교함으로써 그 유효성을 입증하고자 한다.

Ⅱ. 문자열 영상에서의 각 연결 성분간 간격추정

본 논문에서는 각 연결 성분 간의 간격을 계산하고 이를 단어 간 또는 문자 간 간격으로 분류함으로써 개별적인 단어를 추출하는데 주안점을 두고 있다. 이를 위해 먼저 기존의 공간적 거리 측정 방식인 BB, RLEH, CH 방법을 도입하여 그 특성을 분석하였다. 또한 각 측정법을 수정 보완함으로써측정상의 오류를 줄이고 단어 분리 성능을 개선하고자 하였다. 이에 대한 자세한 설명은 다음과 같다.

2.1. BB, RLEH, CH 방법에 의한 간격추정

BB 방법은 그림 1 (a)에 나타낸 바와 같이 서로 인접한 두 연결 성분 간의 간격을 이들을 둘러싸는 최소 사각형 사이의 수평적 직선 거리로 간단히 정의한다. 두 연결 성분이 수평적으로 중첩된 경우에는 그 간격을 0으로 처리한다. RLEH 방식에서는 그림 1 (b)에서와 같이 주어진 두 연결 성분 간의 간격을 추정하기 위해 몇 가지 경험적 기법과 함께 최소 런 거리 또는 최소 유크리디언 거리가 사용된다. 두 연결 성분이 미리 주어진 문턱값(threshold) 이상으로 수직적으로 중첩되면 간격 추정을 위해 런 거리가 사용되며 그렇지 않은 경우에는 유크리디언 거리가 적용된다. 그림 1 (c)에 나타낸 CH 방식에서는 각 연결 성분을 둘러싸는 최소 다각형을 먼저 구하고 인접한 두 다각형의 무게중심을 잇는 직선과 두 다각형이 만나는 두 개의 교점을 계산한다. 최종적으로 두 연결 성분 간의 간격은 이들 두 교점 사이의 유크리디언 거리로써 정의된다.



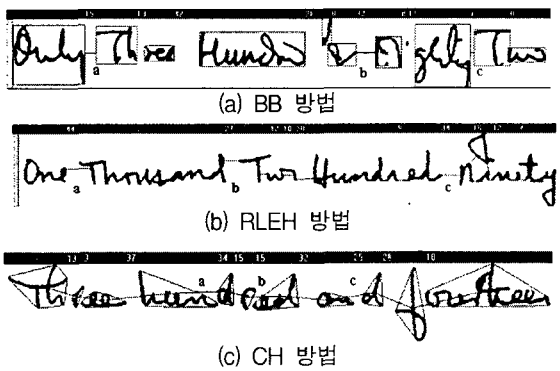
(a) BB 방법

(b) RLEH 방법

(c) CH 방법

그림 1. 연결 성분 간 간격 추정 방법
Fig. 1. Gap estimation using BB, RLEH and CH method

그러나 전술한 세 가지 측정방식은 모두 과대추정 또는 과소추정과 같은 근본적인 문제를 가지고 있다. BB 방법에서는 연결 성분이 수평적으로 돌출한 머리 또는 꼬리 부분을 가지고 있는 경우에 간격 측정 과정에서 과소추정 문제를 발생시킨다(그림 2 (a)의 'a', 'b', 'c' 간격 참조). RLEH 방법의 경우에도 그림 2 (b)에 표시된 'a', 'b', 'c' 간격의 예와 같이 두 연결 성분의 윤곽선 모양 또는 수직적 상호 위치 관계에 따라 과소추정 또는 과대추정의 결과를 낳을 수 있다. CH 방법의 경우에는 그림 2 (c)의 'a', 'b', 'c' 간격에서와 같이 연결 성분의 폭이 상대적으로 넓고 그 시작과 끝 부분에 위로 긴 문자인 어센더(ascender) 또는 그 반대인 디센더(descender)를 포함하고 있는 경우에 최소 다각형을 추정하고 간격을 계산하는 과정에서 과대추정 문제를 빈번히 발생시킬 수 있다.



(a) BB 방법

(b) RLEH 방법

(c) CH 방법

그림 2. 각 측정법 별 측정 오류의 예
Fig. 2. Examples of misestimation

본 논문에서는 각 연결 성분에 대한 최소 사각형 또는 다각형 추정을 위한 새로운 제약 조건과 경험적 기법 등을 도입하여 전술한 이러한 각 측정 방법들의 문제점을 보완하고자 한다.

2.2. 수정된 거리 측정 방법

먼저, 본 논문에서는 수평적으로 돌출된 머리 또는 꼬리 부분에 의해 발생되는 BB 방법에서의 과소추정 문제를 줄이기 위해 각 연결 성분을 둘러싸는 최소 사각형의 좌우 경계선을 조정하고자 한다. 일반적으로 연결 성분의 머리와 꼬리 부분은 그림 3 (a)에 나타낸 바와 같이 수평 성분의 단일 스트로크 형태로 구성되며 연결 성분의 최좌측 지점으로부터 특정 노드(N), 그리고 최우측 지점으로부터 특정 노드(M) 사이의 영역으로 각각 정의된다. 이때 노드 포인트는 전체 영상의 평균 스트로크 두께를 W라고 가정하였을 때 연결 성분의 수직 히스토그램이 βW 보다 커지는 지점으로 설정된다. 연결 성분의 평균 스트로크 두께는 참고 문헌 [6]에서 정의된 방식에 따라 계산된다. 즉, 전체 영상에서 스트로크의 두께가 일정하다고 가정하였을 때 각 연결 성분은 길이 L과 두께 W로 표현할 수 있으며 연결 성분의 둘레 C와 검의 화소의 수 P는 2(L+W)와 LW로 각각 정의된다. P와 C는 문장 금액 영상에 대한 기본적인 전처리 과정에서 자연스럽게 구해지므로 스트로크의 두께 W도 쉽게 계산될 수 있다. 파라메타 β는 잡음과 왜곡 등을 고려하여 경험적으로 1.25로 설정된다.

최종적으로 연결 성분에 대한 최소 사각형의 좌우 경계선은 아래의 식에 의해 조정된다.

$$L_x^{\text{new}} = L_x^{\text{old}} + \alpha H_{\text{wd}} \quad (1)$$

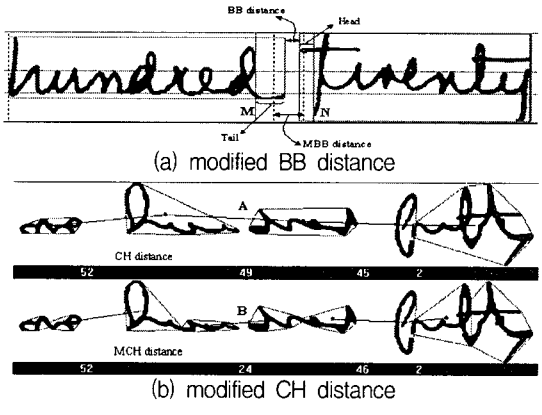
$$R_x^{\text{new}} = R_x^{\text{old}} - \alpha T_{\text{wd}} \quad (2)$$

여기서, H_{wd}와 T_{wd}는 머리와 꼬리 부분의 폭을 각각 나타낸다. 또한 파라메타 α는 이들 돌출 부분이 연결 성분의 몸통 영역에 위치할 경우에 0.35, 그리고 그 외 영역의 경우에는 0.25로 설정된다. 그림 3 (a)의 예에서 볼 수 있듯이 최소 사각형의 좌우 경계선이 적절히 조정됨으로써(MBB) 돌출된 머리 또는 꼬리 부분을

가지는 연결 성분 사이의 간격이 과소추정 되지 않고 비교적 정확히 계산됨을 알 수 있다.

이러한 최소 사각형의 좌우 경계선 조정 개념은 RLEH 방법에 의한 간격 측정에도 유효하게 적용될 수 있다. 수정된 RLEH(MRLEH) 방법에서는 좌우 경계선이 적절히 조정된 최소 사각형 내에 위치하는 연결 성분의 윤곽선만을 간격 측정에 적용함으로써 연결 성분의 머리 및 꼬리 부분에 의한 측정 오류를 줄이고 윤곽선 모양에 따른 측정의 민감도를 다소 둔화 시키고자 하였다.

마지막으로, 본 논문에서는 각 연결 성분을 수직 이등분하여 얻어진 좌우 영역에 대해 각각 개별적인 최소 다각형을 구하고 인접한 두 연결 성분 간의 거리는 좌측 연결 성분의 우측 영역을 둘러싸는 최소 다각형과 우측 연결 성분의 좌측 영역을 둘러싸는 최소 다각형 사이의 거리로써 정의하는 수정된 CH(MCH) 방법을 새로이 제안하였다. 각 연결 성분을 둘러싸는 하나의 최소 다각형에 기반하여 간격을 측정하는 기존의 CH 방법은 그림 3 (b) 의 윗 그림에 나타낸 바와 같이 연결 성분의 폭이 상대적으로 넓고(폭이 높이보다 큰 경우) 그 영역의 시작과 끝부분에 어센더 또는 디센더가 포함되어 있는 경우에 심각한 과대추정 문제를 발생 시킬 수 있으나 제안된 MCH 방식은 아래 그림에서와 같이 이러한 연결 성분을 이등분하고 각 좌우 영역을 해당되는 좌우 연결 성분과의 거리 측정에 독립적으로 적용함으로써 과대추정 문제를 최소화 할 수 있다.



(a) modified BB distance

(b) modified CH distance

그림 3. (a) MBB와 (b) MCH방법에 의한 간격 추정
Fig. 3. Gap estimation

Ⅲ. 단어 분리 실험

본 논문에서는 CENPARMI의 IRIS를 표준 데이터베이스로 사용하여 수정된 각 거리 측정법에 기반한 단어 분리 실험을 수행하고 그 성능을 기존 방식과 비교 분석하였다. IRIS 데이터베이스는 북미 지역의 은행에서 실제로 유통되는 수표로부터 문장 금액을 추출한 것으로서 이진화 과정과 문장 금액 영역의 추출 과정에서 발생되는 상당한 잡음과 모양 왜곡, 그리고 단어간 간격의 심한 불규칙성을 포함하고 있다. 실제 실험 과정에서는 IRIS 데이터베이스에 포함되어 있는 1030개의 영상 샘플에 대해 단어 분리 실험을 수행하고 그 결과를 분석하였다.

3.1. 2-클래스 군집화에 기반한 간격 분류

주어진 문장 금액 영상으로부터 단어를 최종적으로 분리하기 위해서는 주어진 거리 측정법에 의해 추정된 각 연결 성분 사이의 간격을 단어 간 간격(IWG) 또는 문자 간 간격(ICG)으로 분류하는 과정이 추가적으로 수행되어야 한다. 본 논문에서는 LBG 알고리듬[8]을 이용한 2-클래스 군집화 과정을 수행하여 각 간격을 IWG 또는 ICG로 분류함으로써 문장 금액에서 단어를 분리한다.

실험에서는 단어 추출을 위한 군집화 과정을 수행하기 전에 하나의 단어만을 포함하고 있는 영상 즉, 영상내의 모든 간격이 ICG로만 구성된 경우와 영상에 포함된 모든 단어내의 문자 또는 연결 성분이 서로 붙어있는 경우 즉, 모든 간격이 IWG로만 이루어진 영상 샘플들을 아래에 기술한 경험적 기법에 따라 먼저 분류한다

1) 세가지 거리 측정법 별로 영상 내 최대, 최소, 평균 간격을 각각 계산한다.
2) 주어진 입력 영상 내 최좌측 연결 성분에서 최우측 성분까지의 폭이 전체 영상폭의 15% 미만인 경우에 영상 내 모든 간격을 ICG로 간주한다.
3) 그 폭이 전체 영상의 35% 미만이고 두 가지 측정법 별 최대 간격이 전체 데이터베이스로부터 계산한 전체 평균 간격보다 작거나 모든 측정법에서의 평균 간격이 전체 평균 간격의 70%보다 적은 경우에 모든 간격을 ICG로 간주한다.
4) 그 폭이 전체 영상의 35% 이상이고 두 가지 측

정법 별 최소 간격이 전체 평균 간격 보다 큰 경우에 모든 간격을 IWG로 간주한다.

아래의 그림에서는 연결 성분 간의 모든 간격이 ICG 또는 IWG로만 이루어진 영상과 전술한 방식에 의한 분류 결과의 예를 나타내었다.

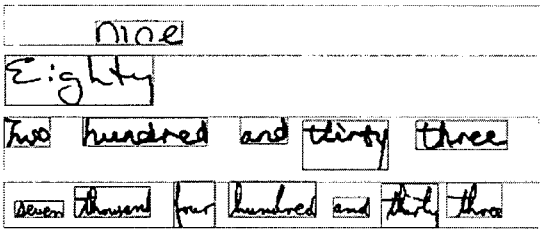


그림 4. 문자 간 간격(ICG) 또는 단어 간 간격(IWG)만을 포함하고 있는 영상 및 그 분류 예
Fig. 4. Examples of image samples containing only inter-character gaps (ICG) or only inter-word gaps (IWG), and their classification

실제로 실험에 사용된 문장 금액 영상 샘플들은 대부분 ICG와 IWG를 동시에 포함하고 있으며 2-클래스 군집화 과정을 통해 이를 분류한다. 표 1에 나타낸 단어 분리 실험 결과로부터 기존의 거리 측정법에서는 RLEH 방법이 70.1%의 단어 분리율(correct separation rate)을 보임으로써 다른 측정법에 비해 조금 더 나은 성능을 나타냄을 알 수 있다. 여기서 분리율은 문장 금액 내의 모든 단어가 성공적으로 분리된 영상 샘플의 수와 전체 데이터베이스와의 비로 정의된다. 또한 수정된 각 거리 측정법은 대응되는 기존의 측정법에 비해 2-3% 정도 향상된 단어 분리 성능을 보임으로써 기존 방법이 가지는 문제점이 어느 정도 보완되었음을 알 수 있다. 특히, MRLEH 방법은 동일한 데이터베이스를 사용한 다른 연구 결과와 비교하더라도 더 나은 실험 결과를 보여준다.

표 1. 각 측정 방식 별 단어 분리 실험 결과
Table 1. Experimental results of word separation

Distance Measures	BB	RLE	CH	MBB	MRLEH	MCH	Zhou [7]	Kim [9]
Separation rate(%)	69.0	70.1	68.3	71.8	72.7	71.7	72.0	70.4
DB (samples)	IRIS (1030)						IRIS (389)	IRIS (100)

그림 5에서는 기존의 측정법에서 발생되는 과소추정 또는 과대추정 문제로 인한 단어 분리 에러가 수정된 거리 측정법에 의해 개선됨을 증명하는 실제 예를 나타내었다. 그림 5 (a)와 (b)에 나타낸 문장 금액에서 "hundred twenty"와 "Thou sand Two"는 기존의 BB와 RLEH 방법에서는 과소추정으로 인해 하나의 단어로 잘못 합쳐져 분류되나 수정된 측정법에 의해 두개의 단어로 정확히 분리된다. 또한 그림 5 (c)는 CH 방법의 과대추정 문제로 인해 세 개의 부분으로 잘못 나누어진 "hundred"가 수정된 방식인 MCH에 의해 하나의 단어로 제대로 합쳐져 분리됨을 보여준다.

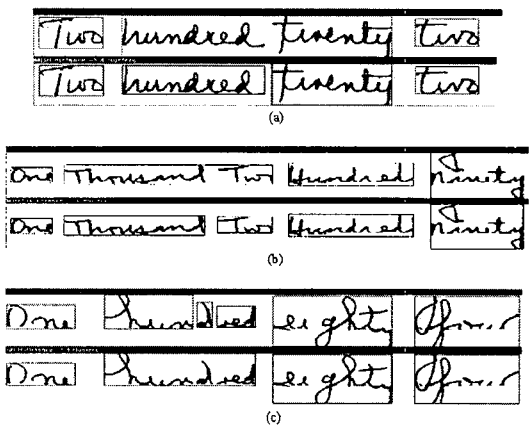


그림 5. 기존 측정 방식에서의 단어 분리 에러(각 그림의 윗 부분)와 수정된 방식에 의한 교정 예 (a) BB (b) RLEH (c) CH 기반의 측정 방법
Fig. 5. Examples showing word separation error by original distance measures and correction by their modified versions (a) BB (b) RLEH and (c) CH method

Ⅳ. 결 론

본 논문에서는 연결 성분 간의 거리 측정에 기반하여 수표 영상 내의 문장 금액에서 단어를 효율적으로 분리하기 위한 방법을 제안하였다. 기존의 일반적인 측정 방식인 BB, RLEH, CH 방법은 과대추정 또는 과소추정과 같은 근본적인 문제를 내포하고 있어 무제약적 필기 스타일에 의한 문자의 왜곡과 필기 공간의 제약으로 인한 연결 성분 간 중첩 및 접촉, 그리고 단어 간 또는 문자 간 간격의 불규칙성이 심하게 나타나는

문장 금액에 적용하는데 많은 어려움을 가진다.

본 논문에서는 기존의 측정 방법들을 수정 보완하여 각 방식의 단어 분리 성능을 향상시키고자 하였다. BB 방법의 경우에는 연결 성분을 둘러싸는 최소 사각형의 좌우 경계선을 머리 및 꼬리 부분의 돌출 정도에 따라 조정함으로써 과소추정 문제를 줄이고자 하였으며 RLEH 방법에 대해서도 좌우 경계선 조정 개념을 적용하여 연결 성분의 모양에 따른 측정의 민감도를 둔화 시키고자 하였다. CH 방법에서는 두 연결 성분 사이의 간격을 수직 이등분 된 좌측 연결 성분의 우측 영역과 우측 연결 성분의 좌측 영역을 각각 둘러싸는 최소 다각형 사이의 거리로 정의함으로써기존 방법이 가지는 과대추정 문제를 최소화하고자 하였다. CENPARMI의 IRIS를 데이터베이스로 사용한 단어 분리 실험에서 수정된 각 거리 측정법이 대응되는 기존의 방법에 비해 2-3% 정도 향상된 단어 분리 성능을 보임을 확인하였다.

향후 연구 과제로는 서로 다른 형태의 거리 측정법들을 결합하여 각 개별 측정법이 가지는 단점을 상호 보완하고 전체 단어 분리 성능을 보다 향상시킬 수 있는 효과적인 결합 방법에 대한 연구를 계속 진행하고자 한다.

참고문헌

[1] 임길택, 진성일, "Karhunen-Loeve 변환 기반의 부분 공간 인식기와 결합된 다중 노벨티 인식기를 이용한 필기체 숫자 인식, "전자공학회 논문지, 제35권 C편 제6호, pp. 88-98, 6월 1998.

[2] A. Ei-Yacoubi, M. Gilloux, R. Sabourin, and C.Y. Suen, "An Hmm-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 8, pp. 752-760, Aug. 1999.

[3] D. Guillevic and C.Y. Suen, "Recognition of Legal Amounts on Bank Cheques," Pattern Analysis and Applications, vol. 1, no. 1, pp. 28-41, 1998.

[4] G. Seni and E. Cohen, "External Word Segmentation of Off-line Handwritten Text Lines," Pattern Recognition, vol. 27, no. 1, pp. 41-52, 1994.

[5] U. Mahadevan and R.C. Nagabushnam, "Gap Metrics for Word Separation in Hand written Lines," Proc. ICDAR, vol. 1, pp. 124-127, 1995.

[6] J. Schürmann, "Document Analysis - from Pixels to Contents," Proc. IEEE, vol. 80, no. 7, pp. 1101-1119, July 1992.

[7] J. Zhou, C.Y. Suen, and K. Liu, "A Feed back-based Approach for Segmenting Hand written Legal Amounts on Bank Cheques," Proc. ICDAR, pp. 887-891, 2001.

[8] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Communications, vol. 28, no. 1, pp. 84-95, Jan. 1980.

[9] K.K. Kim, J.H. Kim, Y.K. Chung, and C.Y. Suen, "Legal Amount Recognition Based on the Segmentation Hypotheses for Bank Check Processing," Proc. ICDAR, pp. 964-967, 2001.

저자약력

김경민(Kyoung-Min Kim)

고려대 전기공학과졸업(1988)전 동대학원 석사, 박사

현재 여수대학교 전자통신전기공학부 부교수

※관심분야 : 컴퓨터 비젼, 퍼지및 신경 회로망 자동화 자동화