퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 1, pp. 57-62

연결 성분 간 간격 측정에 의한 필기체 수표 금액 문장에서의 단어 추출

Word Separation in Handwritten Legal Amounts on Bank Check by Measuring Gap Distance Between Connected Components

김인철*
In-Cheol Kim

*Centre for Pattern Recognition and Machine Intelligence (CENPARMI)
Concordia University, Montreal, Canada

요 약

본 논문에서는 연결 성분 간의 공간적 간격에 기반하여 수표 영상 내의 필기체 문장 금액에서 단어를 효율적으로 추출하기 위한 방법을 제안한다. 인접한 연결 성분 간의 거리 추정을 위한 기존의 방식들은 과대추정 또는 과소추정 문제로 인한 단어 분리 오류를 초래할 수 있으나 본 논문에서는 이러한 문제를 줄이기 위해 각 측정 방식들을 수정 보완하였다. 또한 본 논문에서는 서로 다른 형태의 세 가지 거리 측정법들을 효과적으로 결합하여 각 개별 측정법이 가지는 단점을 상호 보완하고 전체 단어 추출 성능을 좀더 향상시킬 수 있는 4-클래스 군집화에 기반한 결합 방법을 새로이 제안하였다. 문장 금액에 대한 단어 추출 실험 결과로부터 수정된 각 거리 측정법이 대응되는 기존의 측정법에 비해 2-3% 정도 향상된 단어 분리율을 보임을 확인하였다. 또한 제안된 4-클래스 군집화에 기반한 결합 방식은 각 측정 방식에서 개별적으로 발생하는 에러뿐만 아니라 두개의 방식에서 동시에 나타나는 에러도 효율적으로 감소시킴으로써 전체 단어 분리 성능을 향상 시킬 수 있었다.

Abstract

We have proposed an efficient method of word separation in a handwritten legal amount on bank check based on the spatial gaps between the connected components. The previous gap measures all suffer from the inherent problem of underestimation or overestimation that causes a deterioration in separation performance. In order to alleviate such burden, we have developed a modified version of each distance measure. Also, 4 class clustering based method of integrating three different types of distance measures has been proposed to compensate effectively the errors in each measure, whereby further improvement in performance of word separation is expected. 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. In addition, the proposed combining method based on 4-class clustering achieved further improvement by effectively reducing the errors common to two of three distance measures as well as the individual errors.

Key Words : 단어 추출, 연결 성분 간 간격 추정, 군집화에 의한 간격 분류, 필기체 문장 금액

1. 서 론

필기체 문자의 자동 인식은 음성 인식과 더불어 사람과 컴퓨터 사이의 보다 자연스러운 인터페이스 환경을 구현하기 위한 중요한 한 요소로서 그 연구가 국내외적으로 매우 활발히 진행되어 왔다. 일반적인 필기체 문자열 또는 문장에 대한 인식 연구는 필기자의 무제약적 필기 습관에 따른 문자의 심한 왜곡, 어휘 수 문제 등으로 인해 그 수행에 많은 어려움을 가지고 있다. 따라서 현재까지 대부분의 연구는 우편 봉투상의 주소 인식, 수표 상의 문자 금액 인식 등과 같이 제한된 어휘를 가지는 일부 응용 분야에 대한 인식시스템 구현에 주안을 두고 있다 [1-3].

주어진 문자열 영상으로부터 개별적인 단어를 먼저 추출하고 이를 인식한 후에 문맥 정보 등을 이용하여 전체 문자열을 해석하는 것은 필기체 문자열 인식을 위한 가장 일반적인 접근 방식 중의 하나이다. 이러한 방식에 기반하여 견실한 인식 결과를 얻기 위해서는 문자열로부터 개별적인 단어를 정확히 추출할 수 있는 전처리 과정의 개발이 선행되어야 한다. 기존의 관련 연구에서는 문자열을 구성하는 연결 성분(connected component)들 사이의 공간적인 간격(gap)을 특정한 거리 측정법을 이용하여 계산하고 문자 간 간격(inter-character gap)과 단어 간 간격(inter-word gap)으로 분류함으로써 단어를 추출하는 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 거리

접수일자 : 2003년 7월 1일
완료일자 : 2003년 9월 10일
감사의 글 : 이 논문은 한국과학재단의 해외 Post-doc 연수지원에 의하여 연구되었음.

를 계산하는 BB 방법과 연결 성분 간의 최소 유크리디언 거리(minimum Euclidean distance) 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 세 가지 거리 측정법, BB, RLEH, CH는 모두 간단하면서도 효율적으로 연결 성분 간 간격을 측정할 수 있다는 장점을 가지지만 과대추정(overestimation) 또는 과소추정(underestimation)과 같은 근본적인 문제를 내포하고 있어 문자 분리 과정에서 많은 오류를 초래할 수 있다. 본 논문에서는 측정 과정에서의 이러한 오류를 줄이고 단어 분리 성능을 향상시키기 위해 각 측정법들을 먼저 수정 보완하였다. BB 방법의 경우에는 연결 성분을 둘러싸는 최소 사각형의 좌우 경계선을 조정함으로써 수평적으로 돌출된 머리 및 꼬리 부분에 의해 주로 발생되는 과소추정 문제를 줄이고자 하였으며, RLEH 방법에 대해서도 좌우 경계선 조정 개념을 적용하여 연결 성분의 윤곽선으로부터 직접 거리를 측정함으로써 발생되는 측정의 민감도를 둔화 시키고자 하였다. CH 방법에서는 각 연결 성분을 수직 이등분하여 얻어진 좌우 영역에 대해 각각 최소 다각형을 구하고 두 연결 성분 간의 거리는 좌측 연결 성분의 우측 영역을 둘러싸는 최소 다각형과 우측 연결 성분의 좌측 영역을 둘러싸는 최소 다각형 사이의 거리로 정의함으로써 기존의 CH 방법이 가지는 과대추정 문제를 최소화 하고자 하였다. 또한 본 논문에서는 이들 세 가지 측정법을 효과적으로 결합하여 각 개별 측정법이 가지는 단점을 상호 보완하고 전체 단어 추출 성능을 좀더 향상시키기 위해 4-클래스 군집화에 기반한 결합 방법을 새로이 제안하였다.

본 논문에서는 수정된 각 거리 측정법과 새로이 제안된 결합 방식을 수표 상에 표시된 필기체 문장 금액(legal amount)에서의 단어 추출 문제에 적용하여 그 성능을 비교 분석 하였다. 수표 내 문장 금액에서는 일반적인 필기체 문자열이 가지는 필기자의 무제약적 필기 스타일에 의한 필기 문자의 왜곡 외에도 필기 공간의 제약으로 인한 연결 성분간 수평적 중첩(overlapping) 및 접촉(touching), 그리고 단어 간 또는 문자 간 간격의 불규칙성이 심하게 나타난다. 따라서 과대추정 또는 과소추정 문제점을 가지고 있는 기존의 측정 방식으로는 연결 성분 사이의 간격을 정확히 측정하는데 많은 어려움을 가지며 수정된 방식을 적용하여 기존 방식의 문제점을 최소화하고 단어 분리 성능을 향상시킴으로써 그 유효성을 입증하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 기존의 간격 측정법들의 특성 및 문제점들을 분석하고 이를 보완하기 위해 제안된 수정된 간격 측정법에 대해 자세히 설명한다. 3 장에서는 4-클래스 군집화에 기반한 결합 방식에 대한 설명과 각 측정 방식 별 단어 분리 실험 결과 및 고찰을 기술하고 마지막으로 4 장에서 결론을 맺는다.

2. 연결 성분 간 간격 추정을 위한 거리 측정법

본 논문에서는 수표 내 문장 금액 인식을 위한 전처리 단계로서 각 연결 성분 간의 공간적인 간격을 정확히 계산하고 이를 단어 간 또는 문자 간 간격으로 분류함으로써 개별적인 단어를 추출하는데 주안점을 두고 있다. 이를 위해 먼저 기존의 공간적 거리 측정 방식인 BB (bounding box), RLEH(run-length/Euclidean with heuristics), CH(convex hull)를 도입하여 그 특성을 분석하였다. 또한 각 측정법을 수정 보완함으로써 측정상의 오류를 줄이고 단어 분리 성능을 개선하고자 하였다. 이에 대한 자세한 설명은 다음과 같다.

2.1 공간적 거리 측정 방법

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.

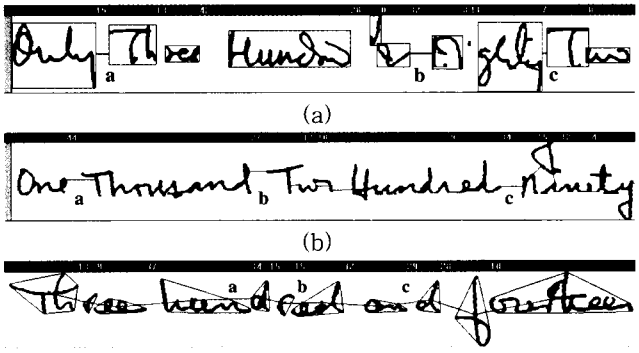

그림 1. 간격 추정 (a) BB, (b) RLEH, (c) CH 방법.
Fig. 1. Gap estimation (a) BB, (b) RLEH, (c) CH method.

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 1 (a)의 'a', 'b', 'c' 간격 참조). RLEH 방법의 경우에는 그림 1 (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. 이러한 BB와 RLEH의 문제점은 CH 방법을 도입함으로써 어느 정도 해결할 수 있다. 그러나 CH 방법에서도 그림 1 (c)의 'a', 'b', 'c' 간격에서와 같이 연결 성분의 폭이 상대적으로 넓고 그 시작과 끝부분에 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. 본 논문에서는 각 연결 성분에 대한 최소 사각형 또는 다각형 추정을 위한 새로운 제약 조건과 경험적 기법 등을 도입하여 전술한 각 측정 방법들의 문제점을 보완하고자 한다.

2.2 수정된 거리 측정 방법

먼저, 본 논문에서는 각 연결 성분을 둘러싸는 최소 사각형의 좌우 경계선을 조정함으로써 수평적으로 돌출된 머리 또는 꼬리 부분에 의해 발생되는 BB 방법에서의 과소추정 문제를 줄이고자 한다. 일반적으로 연결 성분의 머리와 꼬리 부분은 그림 2 (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 'N'과 'M'은 전체 영상의 평균 스트로크 두께를 W라고 가정하였을 때 연결 성분의 수직 히스토그램이 βW 보다 커지는 지점으로 설정된다. 여기서 파라메타 β는 잡음과 왜곡 등을 고려하여 경험적으로 1.25로 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 L×W로 각각 정의된다. P와 C는 문장 금액 영상에 대한 기본적인 전처리 과정에서 자연스럽게 구해지므로 스트로크의 두께 W도 쉽게 계산될 수 있다.

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

$$L_x^{new} = L_x^{old} + \alpha H_{wd} \tag{1}$$

$$R_x^{new} = R_x^{old} + \alpha T_{wd} \tag{2}$$

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 2 (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. 부가적으로, MRLEH 방법에서는 인접한 두 연결 성분이 수직적으로 겹치지 않는 경우에 런 거리 또는 유크리디언 거리를 사용하지 않고 MBB 방법을 적용함으로써 'i' 또는 'j' 와 같은 문자에서 몸통 부분과 점 사이의 간격을 정확히 측정하고 잡음 및 이진화 과정에서 발생되는 점 또는 문자의 깨어진 일부분과 같은 작은 연결 성분으로 인해 발생되는 측정상의 오류를 줄이고자 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 2 (b)의 윗 그림에 나타낸 바와 같이 연결 성분의 폭이 상대적으로 넓고(폭이 높이 보다 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 대응되는 좌우 연결 성분과의 거리 측정에 독립적으로 적용함으로써 과대추정 문제를 최소화 할 수 있다. 또한 MCH 방법의 경우에도 MRLEH 방법에서와 같이 인접한 두 연결 성분이 수직적으로 겹치지 않는 경우에 두 최소사각형 사이의 유크리디언 거리가 아닌 그 수평적 거리로써 간격을 정의한다.

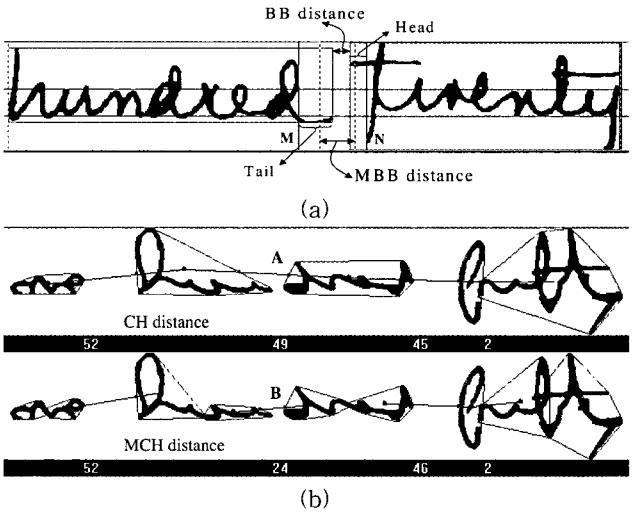

그림 2. (a) MBB와 (b) MCH 방법에 의한 간격 추정.
Fig. 2. Gap estimation using (a) modified BB distance and (b) modified CH distance.

3. 단어 분리 실험 및 고찰

본 논문에서는 CENPARMI의 IRIS[7]를 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 단어 간 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 군집화에 기반한 간격 분류

주어진 문장 금액 영상으로부터 단어를 추출하기 위해서는 각 연결 성분 사이의 간격을 추정된 거리값에 따라 단어 간 간격(inter-word gap, IWG) 또는 문자 간 간격(inter-character gap, ICG)으로 분류할 수 있는 적절한 문턱값이 먼저 정해져야 한다. 그러나 연결 성분 간 간격의 심한 불규칙성으로 인해 문턱값을 적절히 설정하는 것은 쉬운 일이 아니며 일부 연구에서는 경험적 기법에 기반하여 문턱값을 정하는 방법을 제안하고 있다[8]. 본 논문에서는 LBG 알고리즘[9]을 이용한 2-클래스 군집화(clustering) 과정을 수행하여 각 간격을 IWG 또는 ICG로 분류하고 이를 바탕으로 각 연결 성분을 단어 단위로 묶는다.

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) 세 가지 거리 측정법 별로 영상 내 최대, 최소, 평균 간격을 각각 계산한다.
2) 주어진 입력 영상 내 최좌측 연결 성분에서 최우측 성분까지의 폭이 전체 영상 폭의 15% 미만인 경우에 영상 내 모든 간격을 ICG로 간주한다.
3) 그 폭이 전체 영상의 35% 미만이고 두 가지 측정법 별 최대 간격이 전체 데이터베이스로부터 얻어진 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 측정법 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.

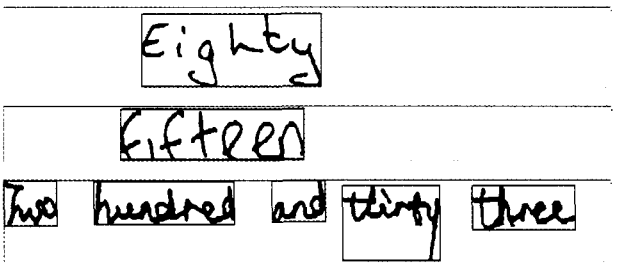

그림 3. 문자 간 간격(ICG) 또는 단어 간 간격(IWG) 만을 포함하고 있는 영상 및 그 분류 예.
Fig. 3. An example of image samples containing only inter-character gaps (ICG) or only inter-word gaps (IWG), and their classification.

실제로 실험에 사용된 문장 금액 영상 샘플들은 대부분 ICG와 IWG를 동시에 포함하고 있으며 2-클래스 군집화 과정을 통해 이를 분류하고 최종적으로 단어를 추출한다. 표 1에 나타낸 실험 결과로부터 기존의 거리 측정법에서는 RLEH 방법이 70.1%의 단어 분리율을 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.

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

Distance Measure	BB	RLEH	CH	MBB	MRLEH	MCH
Correct Separation(%)	69.0	70.1	68.3	71.8	72.7	71.7

그림 4에서는 기존의 측정법에서 발생되는 과소추정 또는 과대추정 문제로 인한 단어 분리 에러가 수정된 거리 측정법에 의해 개선됨을 증명하는 실제 예를 나타내었다. 그림 4 (a)와 (b)에 나타낸 문장 금액에서 "hundred twenty"와 "Thousand Two"는 기존의 BB와 RLEH 방법에서는 과소추정으로 인해 하나의 단어로 합쳐져 잘못 분류되나 수정된 측정법에 의해 두개의 단어로 정확히 분리된다. 또한 그림 4 (c)는 CH 방법의 과대추정 문제로 인해 세 개의 부분으로 잘못 나누어진 "hundred"가 수정된 방식인 MCH에 의해 하나의 단어로 제대로 합쳐져 추출됨을 보여준다.

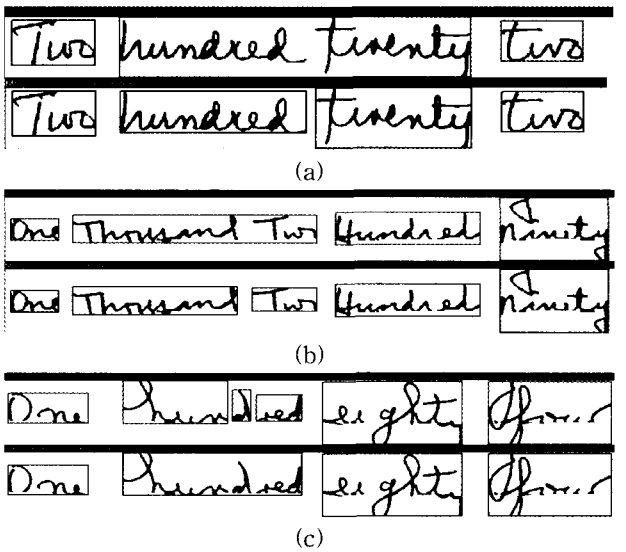

그림 4. 기존 측정 방식에서의 단어 분리 에러(각 그림의 윗부분)와 수정된 방식에 의한 교정 예 (a) BB (b) RLEH (c) CH 기반의 측정 방법.
Fig. 4. Examples showing word separation error by original distance measures (upper part of each figure) and correction by their modified versions (a) BB, (b) RLEH, and (c) CH based method.

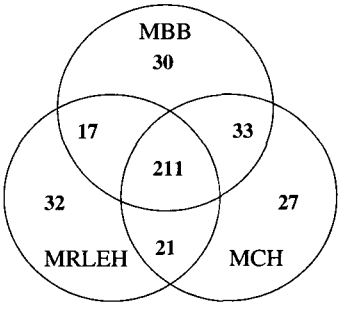

그림 5. 각 거리 측정 방식 별 단어 분리 에러 분포.
Fig. 5. Distribution of separation errors for individual distance measures.

본 논문에서는 전술한 실험 과정에서 서로 다른 방식의 세 가지 거리 측정법에 대해 개별적인 단어 분리 실험을 수행하였으나 이들 세 방법을 통합하여 사용할 경우에 좀더 나은 단어 분리 성능을 얻을 수 있을 것으로 예상된다. 그림 5에서는 각 측정 방식 별 단어 분리 에러의 분포를 벤 다이어그램 형식으로 나타내었다. 많은 에러가 각 측정 방식에서만 개별적으로 나타나지 않고 두 가지 이상의 측정법에서 동시에 발생함을 알 수 있다. 이러한 분포 결과는 다수결 선택 방

식(majority voting)과 같은 단순한 결합 방법으로는 전체적인 성능 향상을 기대할 수 없음을 의미한다. 따라서 본 논문에서는 각 개별 측정법이 가지는 단점을 상호 보완하고 전체 단어 추출 성능을 보다 향상시킬 수 있는 효과적인 결합 방법을 제안하였다.

3.2 4-클래스 군집화에 기반한 세 가지 거리 측정 방법의 결합

본 논문에서는 각 측정 방식에서 개별적으로 발생하는 에러뿐만 아니라 공통적으로 발생하는 에러를 효과적으로 줄이기 위해 아래에 설명된 4 클래스 군집화에 기반한 결합 방식을 제안하였다.

1) 입력 영상 내의 모든 연결 성분 간의 간격에 대해 주어진 거리 측정법을 이용하여 그 거리를 계산하고 LBG 알고리즘에 기반한 군집화 과정을 통해 이들을 4개의 클래스로 분류한다. 이때 각 클래스는 중심값(centroid)의 크기에 따라 정렬화 된다.
2) 군집화의 결과에 따라 각 연결 성분 간 간격에 정수 값 $\alpha \in \{2, 1, -1, -2\}$를 지정한다.

$$\alpha_i = \begin{cases} 2 & \text{for} \quad g_i \in class\,1 \\ 1 & \text{for} \quad g_i \in class\,2 \\ -1 & \text{for} \quad g_i \in class\,3 \\ -2 & \text{for} \quad g_i \in class\,4 \end{cases} \tag{3}$$

여기서, g_i는 주어진 문장 금액 영상내의 i 번째 간격을 의미한다. 결과적으로, 가장 작은 중심값을 가지는 클래스(class 1)에 속하는 간격에는 ICG 가능성을 강조하기 위해 가장 큰 양의 정수 값($\alpha = 2$)이 지정되며 그 반대의 경우에는 가장 작은 음의 값($\alpha = -2$)이 지정되어 IWG 가능성을 강조한다.
3) MBB, MRLEH, MCH 방법에 대해 윗 과정 1), 2)를 반복 수행함으로써 최종적으로 i 번째 간격에 대해 세 가지 정수 값, α_i^{BB}, α_i^{RLE}, α_i^{CH}이 지정된다.
4) 세 정수의 합, $\alpha_i^{BB} + \alpha_i^{RLE} + \alpha_i^{CH}$이 양의 값이면 i 번째 간격을 ICG로 정의한다. 그렇지 않은 경우에 IWG로 정의한다.

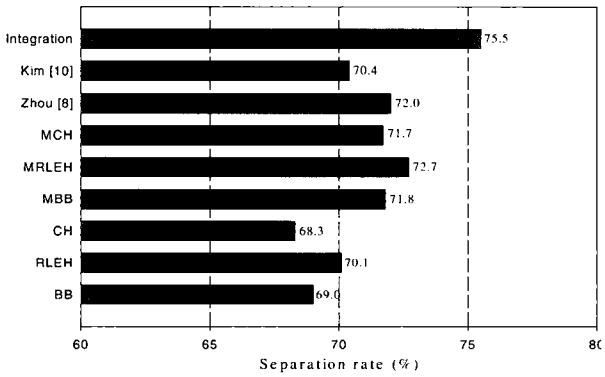

그림 6. 각 방식 별 단어 분리 성능 비교.
Fig. 6. Comparison of word separation rate according to methodologies.

본 논문에서는 전술한 결합 방식을 이용하여 추가적인 단어 분리 실험을 수행하였으며 그림 6에 나타낸 바와 같이 개별적인 거리 측정 방식뿐만 아니라 동일한 데이터베이스를 사용한 다른 연구 결과와 비교하더라도 훨씬 더 높은 성능인 75.5%의 단어 분리율을 얻을 수 있었다.

이러한 성능 향상은 그림 7의 벤 다이어그램을 통해 나타낸 에러 분포를 통해서도 확인할 수 있다. 그림 5에 나타낸 각 개별 측정 방식에 대한 에러 분포와 비교할 때, 각 방식에서 개별적으로 발생하는 에러뿐만 아니라 두 개의 방식에서 동시에 나타나는 에러도 급격히 감소되며 이로 인해 전체적인 성능이 향상되었다.

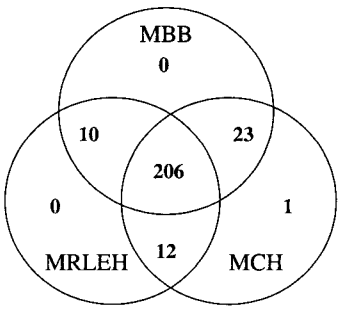

그림 7. 4-클래스 군집화에 기반한 결합 방식에 대한 단어 분리 에러 분포.
Fig. 7. Error distribution by integrated method using 4-class clustering.

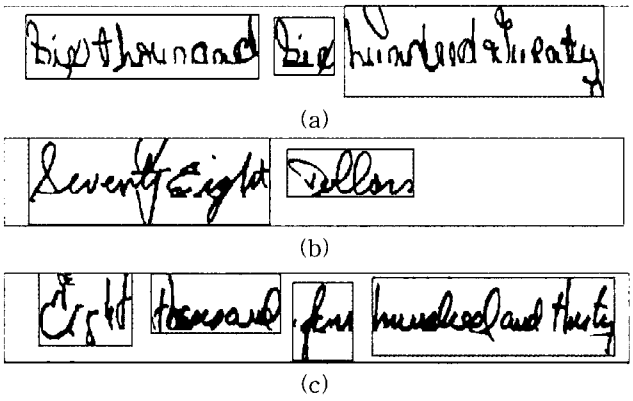

그림 8. 4-클래스 군집화 기반의 결합 방식에 대한 단어 분리 에러 예.
Fig. 8. Examples of word separation error by 4-class clustering based combining method.

그러나 세 가지 방식에서 동시에 발생하는 에러의 경우에는 제안된 결합 방식을 적용하더라도 그 수가 크게 감소하지 않는다. 이러한 결과는 공간적 거리 측정 방식을 적용하기 위한 전제 조건인 단어 간 간격이 문자 간 간격에 비해 일반적으로 더 크게 나타난다는 가정이 수표 내 문장 금액에는 적절하게 맞지 않음을 의미한다. 즉, 필기자의 무제약 필기 스타일과 제한된 필기 공간으로 인해 단어 간 간격이 문자 간 간격에 비해 항상 크게 나타나지는 않으며 인접한 연결 성분 간에 수평적 겹침 또는 접촉이 빈번히 발생한다. 이러한 문제로 인한 단어 추출의 어려움은 그림 8 (a)와 (b)에 나타낸 에러 예로써 잘 설명된다. 또한 필기체 문장 금액에서는 단어 간 간격들 사이의 편차가 문자 간 간격에 비해 상당히 더 크게 나타나는데, 이러한 단어 간 간격의 심한 편차는 그림 8 (c)의 예와 같이 또 다른 형태의 에러를 발생 시킬 수 있다. 즉, 주어진 영상 내의 모든 단어 간 간격이 문자 간 간격에 비해 더 크게 측정되더라도 그 편차로 인해 상대적으로

작은 크기의 단어 간 간격이 군집화 과정에서 문자 간 간격으로 잘못 분류될 수 있다.

전술한 연결 성분 간 간격의 불규칙성 또는 중첩으로 인한 에러를 줄이기 위해서는 공간적 거리 측정 방식 외에 문장 금액을 구성하는데 필요한 단어의 수 또는 문맥 정보 등과 같은 사전 정보를 단어 분리 과정에 도입하는 것이 필요하다. 또한 암시적(implicit) 분리 방법 또는 인식 기반의 분리 기법과 같은 다른 접근 방식을 기존 방식과 결합하여 사용하는 경우에 보다 나은 단어 분리 성능을 얻을 수 있을 것으로 예상된다.

4. 결 론

본 논문에서는 연결 성분 간의 거리 측정에 기반하여 수표 영상 내의 문장 금액에서 단어를 효율적으로 분리하는 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 측정 과정에서의 오류를 줄이고 단어 분리 성능을 향상시키기 위해 기존의 측정법들을 먼저 수정 보완하였다. 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 윤곽선 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 정의함으로써 기존 방법이 가지는 과대추정 문제를 최소화하고자 하였다. 또한 본 논문에서는 이들 세 가지 측정법을 결합하여 각 개별 측정법이 가지는 단점을 상호 보완하고 전체 단어 추출 성능을 보다 향상시킬 수 있는 4-클래스 군집화에 기반한 결합 방법을 새로이 제안하였다.

CENPARMI의 IRIS를 데이터베이스로 사용한 단어 분리 실험에서 수정된 각 거리 측정법이 대응되는 기존의 방법에 비해 2-3% 정도 향상된 단어 분리 성능을 보임을 확인하였다. 또한 제안된 4-클래스 군집화에 기반한 결합 방식은 각 측정 방식에서 개별적으로 발생하는 에러뿐만 아니라 두 개의 방식에서 동시에 나타나는 에러도 효과적으로 감소시킴으로써 추가적인 단어 분리 성능의 향상을 얻을 수 있었다.

향후 연구 과제로는 문장 금액에 관련된 사전 정보를 단어 분리 과정에 도입하거나 암시적 또는 인식 기반의 분리 기법과 같은 다른 접근 방식을 기존 방식과 결합하여 연결 성분 간 간격의 불규칙성 및 중첩으로 인한 에러를 효과적으로 줄일 수 있는 연구를 계속 진행하고자 한다.

참고문헌

[1] D. D'Amato, E. Kuebert, and A. Lawson, "Results from a Performance evaluation of Handwritten Address Recognition Systems for the United States Postal Service," Proc. Int'l Workshop on Frontiers in Handwriting Recognition, pp. 189-198, 2000.

[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 Handwritten Lines," Proc. Int'l Conf. Document Analysis and Recognition, vol. 1, pp. 124-127, 1995.

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

[7] D. Guillevic, "Unconstrained Handwriting Recognition Applied to the Recognition of Bank Cheques," Ph. D Thesis, Concordia University, Montreal, Canada, 1995.

[8] J. Zhou, C.Y. Suen, and K. Liu, "A Feedback-based Approach for Segmenting Handwritten Legal Amounts on Bank Cheques," Proc. Int'l Conf. Document Analysis and Recognition, pp. 887-891, 2001.

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

[10] 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. Int'l Conf. Document Analysis and Recognition, pp. 964-967, 2001.

저 자 소 개

김인철(In-Cheol Kim)

1989년 2월 경북대 전자공학과 졸업(공학사)
1991년 2월 경북대 대학원 전자공학과 졸업(공학석사)
2001년 2월 경북대 대학원 전자공학과 졸업(공학박사)
1991년~1996년 (주) 카스 기술개발실 선임연구원
2001년 6월~2002년 5월 경북대 BK21 정보기술사업단 박사후 연구원
2002년 5월~현재 CENPARMI, Concordia University, Canada 박사 후 연구원

관심 분야 : 형태 인식, Multimodal HCI, 컴퓨터비전, 신경회로망 등

E-mail : kiminc@cenparmi.concordia.ca