온라인 영어 필기의 전역적 특징을 이용한 사전으로부터의 후보단어 선택

최 동 원○ 권 재 욱 김 진 형
한국과학기술원 전산학과

Lexicon Filtering with Global features of On-line English Handwriting

Choi Dong Won Kwon Jae Ook Kim Jin Hyung
Dept. of Computer Science, KAIST

요 약

많은 영어 필기 인식 시스템에서는 인식시에 사용하는 외부 지식으로 단어 사전을 사용한다. 이 단어 사전은 인식시에 생성되는 후보의 범위를 결정하는 것으로, 사전의 크기는 인식 성능에 많은 영향을 주게 된다. 즉, 사전의 크기가 작을 수록 인식 성능이 향상된다. 따라서, 본 논문에서는 온라인 영어 필기의 전체적 모양에 관한 정보라 할 수 있는 전역적 특징을 이용하여 사전의 크기를 줄이고자 하였다. 영어 필기에서 찾아볼 수 있는 전역적 특징으로는 점, 수평획, 하향획의 수, 그리고 하향획의 유형이 있을 수 있는데, 특징 분석과 인공 신경망을 이용하여 추출된다. 사전 필터링은 각각의 전역적 특징의 유무, 또는 단어의 특징 포함 여부 등의 필터링 조건에 따라 수행된다. 모든 전역적 특징을 이용한 사전 필터링 실험의 경우, 약 3%의 오류를 허용하여 사전의 크기를 처음 사전의 1%로 줄일 수 있었다.

1. 서론

지금까지 시도된 많은 문자인식 방법들은 사람의 글자 인식 방법과 무관한 나름대로의 방법에 기초하였다. 그러나, 정확성에 있어서는 아직까지 사람의 인식 능력보다 많이 뒤쳐지므로, 여러 심리학적인 연구에서 밝힌 사람의 문자 인식 체계를 모방한 인식 방법이 고려되어야 할 것이다. 본 논문에서는 사람의 문자 인식 체계 중 한 가지로 밝혀지고 있는 필기의 전체적인 모양 정보에 의한 후보 단어 선택 단계를 모방하여 인식 시스템의 단어 사전 크기의 감소 문제를 다루고자 한다.

사전 필터링은 필기 데이타로부터 추출된 특징을 이용하여 보다 작은 크기의 사전을 구성하는 것이다. 이를 위해서는 필터링에 사용될 특징의 추출 과정이 필요하게 된다. 본 논문에서는 기존의 연구에서 사용하여 온 전역적 특징에 대한 체계적인 분석을 통한 보다 안정된 특징 추출 방법을 제안한다. 또, 추출된 특징을 이용한 효과적인 사전 필터링 방법도 제시하고자 한다.

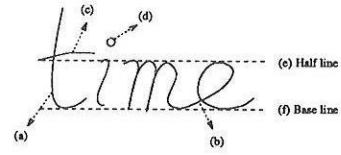
2. 배경

2.1 전역적 특징

영어 필기에서의 전역적 특징이란 필기에서의 획 간의 위치 관계나 각 획의 유형, 또는 획수나 윤곽선의 형태 등을 말한다. 전역적 특징은 지역적 특징에 비해 보다 상위수준의 특징으로서 더욱 많은 정보를 나타내지만, 입력으로 들어온 필기 데이타로부터 바로 얻어질 수 있는 것은 아니다. 그림 1의 경우에는 전역적 특징으로 하향획의 수 6개, 점의 수 1개, 수평획의 수 1개, 폐곡선의 수 1개를 찾아볼 수 있다.

2.2 사전 필터링

사전 필터링이라고 하는 것은 특정한 기준에 따라 사전 내의 각 단어를 취하거나 버리거나 한 후 남은 단어들만을 모아 보다 작은 크기의 사전을 구성하는 것이다. 그림 2에서 'i'가 포함된 단어만을 선택하



(a) 하향획 (b) 상향획 (c) 수평획 (d) 점 (e)(f) 기준선

그림 1: 무제약 필기

는 사전 필터링의 예를 보이고 있다. 사전 필터링이란 다음과 같은 식으로 정형화해볼 수 있다.

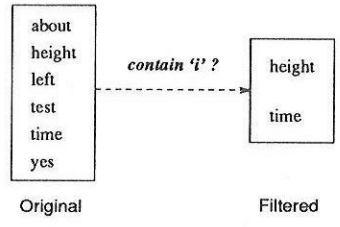


그림 2: 사전 필터링

$\mathcal{L}$이 원래의 사전이고, $\mathcal{L}'$이 필터링된 사전이라 할 때, 필터링 조건 P가 주어지면,

$$\mathcal{L}' = \{w \mid \forall_w \in \mathcal{L},\ P_i(w) = True\}$$

이 된다.

온라인 영어 필기 인식 시스템에서 사용하는 단어 사전을 필터링하려면, 우선 입력된 문자 데이타로부터 필터링을 위한 특징을 추출해야 한다. 이 때 사용하는 특징으로는 필기에서의 전역적 특징을 들 수 있다. 점의 수, 수평획의 수, 하향획의 수 등이 사전 필터링시 사

용될 수 있는 전역적 특징이 된다. 사전 필터링에서는 여러가지 필터링 조건을 다음과 같이 조합하여 사용할 수 있다.

$P_c \longrightarrow (P_c \vee P_c) \mid (P_c \wedge P_c) \mid P$

$P \longrightarrow P_1 \mid P_2 \mid P_3 \mid \cdots \mid P_n$

이러한 조건을 이용하여 필터링된 사전이 인식기에서 사용된다.

사전 필터링을 포함하는 인식 시스템의 구성은 그림 3와 같다.

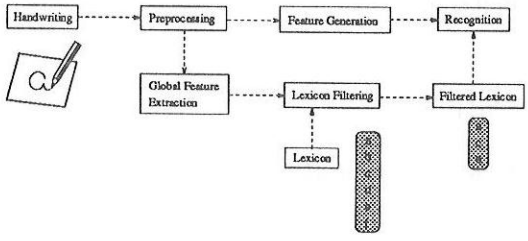


그림 3: 사전 필터링을 이용한 인식 시스템

3. 필터링 방법

3.1 점 및 수평획 정보 추출

필기된 단어 내에서 점의 유무는 'i'나 'j'의 출현 유무를 판단하는 단서를 제공하며, 수평획의 유무는 't', 'f', 또는 'z'의 존재 여부를 예측할 수 있는 단서가 된다.

점은 모든 획을 연결하여 필기하는 경우에도 하나의 독립된 획으로 필기된다는 특징이 있다. 즉, 점을 구성하는 단위는 펜이 획을 그리기 시작한 위치로부터 끝난 위치까지로 한다. 'i'나 'j'의 점은 아래부분에 있는 획의 위쪽에 놓인다는 특징을 이용하여 추출하였다.

수평획의 경우도 점의 경우와 같이 하나의 독립된 획으로 이루어진다. 수평획으로 판단하기 위한 조건으로, 우선 고려중인 획이 다른 하향획과 교차하는가를 조사한다. 예를 들어 설명하면, 그림 4의 't'에서 수평획 AB가 획 CD와 교차함을 볼 수 있다. 여기서 수평획과 교차하는 하향획을 주획이라 하자. 그림 4에서는 획 CD가 고려중인 수평획 AB에 대한 주획이 된다. 고려중인 획이 수평획인지의 여부는 고려중인 획과 주획의 위치와 모양을 고려하여 결정하는데, 다음과 같은 특징을 이용한다.

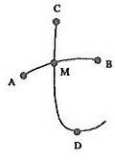


그림 4: 수평획 정보 추출

- 수평획의 기울기 – $\overline{AB}$의 기울기
- 수평획과 주획의 길이비 – $|\widetilde{AB}| : |\widetilde{CD}|$
- 수평획과 주획의 선형성(linearity) – $\frac{|\widetilde{AB}|}{|\overline{AB}|}, \frac{|\widetilde{CD}|}{|\overline{CD}|}$
- 교차점을 중심으로 수평획의 좌우 길이 비 – $|\widetilde{AM}| : |\widetilde{MB}|$
- 교차점을 중심으로 주획의 좌우 길이 비 – $|\widetilde{CM}| : |\widetilde{MD}|$

3.2 하향획의 수를 이용한 필터링

하향획이란 영어 필기시 위에서 아래 방향으로 필기되는 모든 획을 말하는데, 필기 단어의 주된 모양을 결정하는 것으로 윗방향 획에 비하여 규칙적으로 나타난다. 그림 5는 서로 다른 필기 형태에서의 단어 '*nice*'의 획수 변화를 보이고 있다. (a)의 경우 하향획 5개, 상향획 1개가 나타난다. 그러나, (b)의 경우 하향획의 수는 (a)의 경우와 같이 5개이지만, 상향획이 9개나 나타남을 알 수 있다. 즉, 상향획은 글씨를 쓰기 시작하는 시기와 글씨가 끝날 때 펜을 드는 시기, 그리고 글자와 글자간의 연결획(ligature) 등에서 무작위적으로 발생하므로 변화가 많은 반면, 하향획은 대체로 일정하게 나타나며 약간의 변화만을 보인다. 따라서, 하향획의 수는 필기된 단어의 길이에 관한 정보를 비교적 일관되게 제공한다고 할 수 있다.

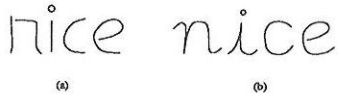


그림 5: 필기 유형에 따른 하향획 수의 변화

추출된 하향획의 수를 이용하여 사전을 필터링을 위해서는 사전에 있는 각 단어의 하향획 수를 알아야 한다. 단어의 하향획 수는 미리 정의된 각 글자의 하향획 수의 변화의 조합으로 구하게 된다. 예를 들어 설명하면, 단어 'by'의 하향획 수의 범위는, 'b'의 획수 변화 범위가 [1,2]이고 'y'의 획수 변화 범위가 [2,2]이므로, [3,4]가 된다.

3.3 하향획의 유형에 의한 필터링

3.3.1 하향획의 유형 분류

영어 필기의 주된 모양을 결정하는 하향획은 다음과 같이 기준선을 중심으로 4가지의 유형으로 분류할 수 있다.

a-type 'a'와 같이 획이 중앙에 위치

l-type 'l'과 같이 획이 중앙과 위쪽에 위치

p-type 'p'의 긴 획과 같이 획이 중앙과 아래쪽에 위치

f-type 필기체 'f'와 같이 획이 위에서 아래에 걸쳐 위치

하향획의 유형에 관한 정보는 필기된 단어의 개략적인 모양에 관한 정보를 제공하므로, 입력된 필기에 대한 예측에 많은 도움을 준다. 그러나, 하향획의 유형 분류는 필기 형식의 다양성으로 말미암아 매우 어려운 문제이다. 하향획의 유형을 분류하기 위한 방법으로 두 가지를 생각할 수 있다. 우선 기준선을 찾은 후 기준선과 각 획의 위치를 비교하여 유형을 분류하는 방법이 있을 수 있다. 하지만 이 방법은 기준선을 찾아야 하는 어려움과 각 하향획의 기준선과의 위치 비교시 발생하는 모호성으로 인하여 현실적으로는 적합하지 못하다.

또 다른 방법으로, 인공 신경망을 이용한 하향획 유형 분류를 생각할 수 있다. 본 논문에서는 2-layer perceptron을 이용하여 하향획의 유형을 분류하였는데, 구조는 그림 6과 같다.

다음은 인공 신경망의 입력으로 사용되는 각 획의 특징 성분이다.

- 획의 가장 윗점의 y 좌표
- 획의 가장 아랫점의 y 좌표
- 획의 선형성(linearity)

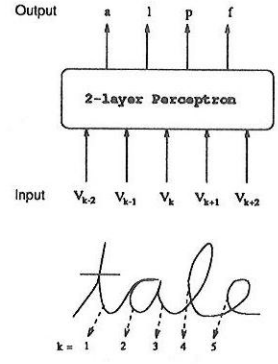


그림 6: 하향획 유형 분류를 위한 인공 신경망 구조

- 획의 가장 윗점에서의 cusp
- 획의 가장 아랫점에서의 cusp

양 옆 두개씩의 획을 포함하여 모두 다섯개의 획에 대하여 위의 특징 성분을 구한 5차 벡터 5개, 즉 25차 벡터가 인공 신경망의 입력으로 사용된다. 이렇게 입력된 정보를 이용하여 인공 신경망은 4가지 획 유형에 대한 확률값을 결과로 출력하게 된다.

3.3.2 단어의 레이블을 이용한 사전 필터링 방법

입력된 필기 단어에 대한 레이블은 하향획의 유형 분류를 거쳐 얻게 된다. 하향획의 유형들은 각각 'a', 'l', 'p', 'f'의 4가지 레이블을 나타내게 된다. 그림 7는 입력으로 들어온 필기 단어에 대한 단어 레이블링의 예이다.

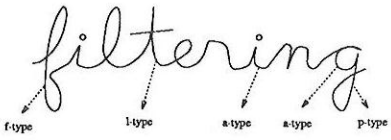


그림 7: 필기 단어의 레이블링 예

영어 단어의 필기는 필기자의 필기 습관에 따라 획순과 획수의 차이가 발생한다. 필터링을 위한 단어 레이블 비교시 이러한 변화는 문제가 된다. 따라서, 사전 필터링을 위하여 단어 레이블을 비교하기 위해서는 이러한 문제를 해결하여야 한다. 우선 단어 필기시 획순의 변화를 해결할 방법을 생각해보자.

일반적인 영어 필기의 획순의 변화는 특정 몇 글자를 필기할 때만 발생하게 된다. 같은 유형의 하향획이 순서가 바뀌어 필기되는 경우는 단어의 레이블의 구성에 아무런 영향이 없다. 또, 일반적인 필기에서는 'x'를 제외하면 한 글자를 필기하는 도중에 다른 글자를 필기하는 경우도 없다. 따라서, 본 논문에서는 글자 내의 서로 다른 유형의 획간에 필기 순서가 바뀌는 경우만 고려한다. 글자 내에서 하향획의 필기 순서가 바뀔 수 있는 글자로는 'b', 'd', 'p', 그리고 'q'가 있다.

'x'의 경우는 단어를 흘림체로 필기하는 경우에 필기 순서에 변화가 발생한다. 그림 8에서 필기 유형에 따른 단어 내에서의 'x'의 획순의 변화를 보이고 있다. (a)의 경우에는 'x'의 두 획이 연속적으로 필기되는 반면, (b)의 경우에는 'x'의 한 획이 제일 마지막에 필기됨을 볼 수 있다. 이러한 획순의 변화를 바로 잡을 수 있어야 단어에 대한 정확한 레이블을 얻을 수 있고, 정확한 단어 레이블을 얻어야만 레이블을 이용한 사전 필터링시 안정된 결과를 얻을 수 있다. 있는지 알아보자.

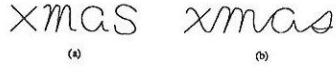


그림 8: 획순 변화의 예

단어 레이블에서의 각 하향획의 레이블 위치는 입력에서 그 하향획의 기하학적인 위치에 따른다. 즉, 필기에서 하향획의 순서를 정렬하기 위해서는 모든 하향획의 기하학적인 정보를 구한 후, X축에 대한 순서 정렬을 한다. 모든 하향획의 순서가 결정되면, 그 순서를 이용하여 단어의 레이블링을 하게 된다.

영어 단어 필기는 필기자의 필기 습관에 따라 여러가지 다른 모양을 갖게 된다. 이렇게 다양한 형태의 필기에 대해서 사전을 필터링하기 위해서는 단어 사전에 있는 각각의 단어의 레이블의 형태가 필기의 변화를 처리할 수 있도록 구성되어야 한다. 본 논문에서는 사전상의 각 단어에 대해 필기 가능한 형태의 레이블을 모두 생성하여 필기된 단어의 레이블과 비교하는 방법을 이용하였다. 우선, 영어 알파벳의 각 글자에 대해서 일반적인 영어 필기시에 나타나는 필기 형태에 따라 글자의 레이블을 결정한다. 그리고, 미리 정의된 글자 레이블의 조합으로 사전상의 단어에 대한 레이블을 구성한다. 다음 예를 통하여 레이블의 생성 과정을 살펴보자.

$$\begin{aligned} \text{Label}[by] &= \text{Label}[b] \cdot \text{Label}[y] \\ &= \{l, la\} \cdot \{ap\} \\ &= \{lap,\ laap\} \end{aligned}$$

이렇게 각 단어에 대하여 가능한 레이블을 생성한 후, 입력으로 들어온 필기 단어의 레이블과 비교하여 사전을 필터링한다. 즉, 입력된 필기 단어의 레이블을 l이라 하고, 사전의 i번째 단어에 대하여 생성된 단어 레이블의 집합을 L_i라 할 때, 다음 식을 통하여 필터링된 사전을 구하게 된다.

$\mathcal{L}$이 원래의 사전이고, $\mathcal{L}'$이 필터링된 사전이라 할 때,

$$\mathcal{L}' = \{w_i \mid \forall_{w_i} \in \mathcal{L},\ l \in L_i,\ L_i = Label[w_i]\}$$

4. 실험 및 결과

본 논문에서 제안한 사전 필터링 방법의 유용성을 보이기 위하여 12명의 필기자가 전자펜으로 자유롭게 필기한 영어단어를 사용하여 실험하였다. 실험은 추출된 특징에 의한 사전 필터링 효과 및 정확성에 관한 실험, 그리고, 필터링된 사전이 실제 인식기의 성능에 미치는 영향을 알아보는 실험을 수행하였다. 사전 필터링에서는 20,000단어 크기의 사전을 사용하였다.

4.1 사전 필터링 실험

4.1.1 점 및 수평획에 의한 필터링 실험

점 및 수평획 정보에 의한 사전 필터링 실험에서의 오류율은 특징 추출시의 오류에서 기인한다. 그림 9에서 보는 바와 같이 점 및 수평획 정보를 이용한 필터링은 큰 효과는 없었다. 하지만, 점과 수평획 정보의 안정된 추출 특성으로 필터링시 감수해야 하는 오류는 상당히 작음을 알 수 있다.

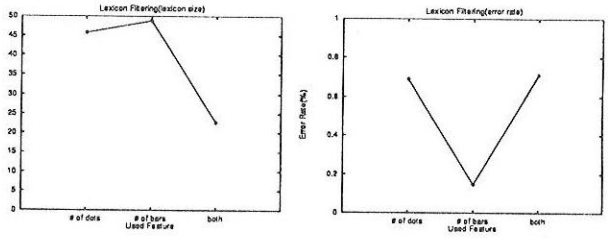


그림 9: 점 및 수평획에 의한 사전 필터링 실험 결과

4.1.2 하향획의 수에 의한 필터링 실험

하향획의 수 정보를 이용한 필터링은 예상했던 대로 상당히 안정된 결과를 보였다. 기본적으로는 필기된 단어 내에서의 하향획의 수는 각 글자의 하향획 수의 조합으로 결정된다. 하지만, 단어 필기의 경우에는 필기를 시작할 때, 필기를 마칠 때, 또는 글자간의 연결획 등에서 또 다른 하향획이 발생할 수 있다. 따라서, 단어의 하향획의 수의 변화를 정할 때, 단순히 글자의 획수 변화의 조합에 의해서 결정을 하기보다는 단어 단위의 획수의 변화를 고려할 필요가 생긴다.

본 논문에서는 단어 내의 글자의 획수 변화의 조합에 의한 단어의 획수 변화에, 단어 단위의 획수의 변화 정도를 반영하였다. 우선, 글자의 획수 변화의 조합에 의한 단어의 획수 변화 범위는 '[Lower,Upper]'로 표현한다. 그리고, 획이 하나 더 첨가되어

서 필기된 경우와 한 획 덜 필기된 경우를 위하여 각각 '[Lower-1,Upper]', '[Lower,Upper+1]'를 고려하여 보았다. 마지막으로 두 가지의 변화를 모두 처리하기 위하여 '[Lower-1,Upper+1]'의 변위를 고려하였다. 그림 10에서 알 수 있듯이 단어 필기시의 획수 변화를 보다 완화시킨 경우에 많은 오류율의 감소를 얻을 수 있었다.

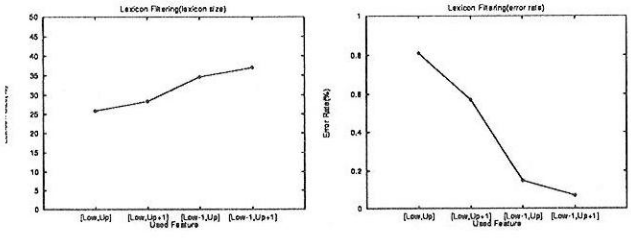


그림 10: 하향획의 수에 의한 사전 필터링 실험 결과

4.1.3 하향획의 유형에 의한 필터링 실험

하향획의 유형을 이용한 사전 필터링은 우선 필기된 단어의 각 획의 유형 분류 결과에 의하여 단어의 레이블을 구성하는 것으로부터 시작된다. 표 1에서 알 수 있듯이, 1등 후보만을 이용한 경우에는 필터링의 효율은 좋지만, 획 유형 분류시의 오류에 따른 많은 필터링 오류가 나타나고 있다. 하지만, 2등 후보를 이용한 경우에는 필터링 효율이 크게 저하되지 않으면서 많은 오류의 감소를 얻을 수 있었다. 또한, 필터링시의 레이블 비교시 입력된 필기의 레이블에서 하나의 획 유형이 잘못된 경우를 고려하여 조건을 완화시킨 경우에는 오류율을 더욱 줄일 수 있었다.

표 1: 하향획의 유형을 이용한 사전 필터링 실험 결과

	사전의 크기(%)	오류율(%)
1등 후보만 이용	0.43	15.85
2등 후보도 이용	0.755	4.58
필터링 조건 완화	1.86	3.11

4.2 필터링된 사전을 이용한 인식 실험

필터링 된 사전을 이용한 인식 실험에서는 5명이 자유롭게 필기한 영어 단어 600개를 이용하였다. 사전 필터링을 위한 전역적 특징으로 점과 수평획의 수, 그리고 하향획의 유형에 의한 단어 레이블을 모두 조합하여 사용하였고, 단어 레이블 비교시에는 완화된 필터링 조건을 사용하였다. 입력된 필기 데이타는 필터링된 사전과 함께 HMM을 이용한 인식 시스템을 통하여 인식이 된다. 그림 11에 필터링된 사전을 이용한 인식 실험 결과가 나타나 있다.

그림 11에 나타난 바와 같이 전체적인 인식률은 향상 되었지만, 필기자에 따라 인식률이 감소하는 경우도 있음을 알 수 있다. 이러한 결과에 대한 원인으로는 수평획이나 점의 필기가 분명하여 추출이 잘 되는 경우에는 사전 필터링의 오류가 적어 전체적인 인식 시스템의 성능 향상이 있었지만, 특징 추출이 잘 되지 않는 필기 유형에 대해서는 필터링에서의 오류로 인식 시스템의 성능이 떨어진 것으로 판단된다. 또, 많은 경우에 있어서의 필터링 오류는 단어의 레이블을 구성하는 하향획의 유형을 분류하는 과정에서 발생한다. 특히, 인공 신경망을 훈련시키는데 사용된 하향획 데이타와 다른 유형의 필기 단어에 대해서 획 유형 분류시 많은 오류가 발생하였다. 이는, 보다 많은 데이타를 통한 인공 신경망의 학습을 통해서 개선될 것으로 생각된다.

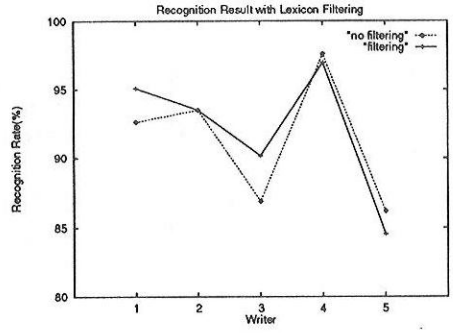


그림 11: 사전 필터링을 이용한 인식 실험 결과

5. 결론

본 논문에서는 사전 필터링을 위하여 영어 필기에서의 전역적 특징을 이용하였다. 사전 필터링에 사용되는 전역적 특징 추출을 위하여, 전자펜으로 입력된 필기 데이타에 대한 통계적인 분석과 인공 신경망을 이용한 효과적인 획 유형 분류 방법을 제안하였다. 본 논문에서 제안된 방법을 통하여 상당히 안정된 사전 필터링 결과를 얻을 수 있었고, 필터링된 사전을 이용한 영어 필기 단어 인식 실험에서 사전 필터링의 효과를 확인할 수 있었다.

사전 필터링 사용시 고려할 점은, 필터링 효율과 필터링시 발생하는 오류간의 절충 과정이다. 실험 결과에서 알 수 있듯이, 필터링 효율을 높이고자 할 경우 오류율도 함께 증가한다. 필터링된 사전의 크기가 작을 수록 인식 성능이 향상된다는 것은 당연한 사실이다. 하지만, 사전을 필터링하는 과정에서 오류가 발생하여 입력된 필기에 해당하는 단어가 제외된다면, 인식기의 성능이 아무리 우수해도 주어진 필기에 대한 올바른 인식이 불가능해지는 것이다. 따라서, 사전 필터링을 이용하고자 하는 인식 시스템의 성능에 따라 사전 필터링 방법을 선택해야 할 것이다.

참고 문헌

[1] Michel Gilloux, Jean-Michel Bertille and Manuel Leroux, "Recognition of Handwritten Words in a Limited Dynamic Vocabulary," *IWFHR-3, Buffalo New York, USA*, pp 417–422, May, 1993.

[2] F. Kimura, M. Shridhar and N. Narasimhamurthi, "Lexicon Directed Segmentation - Recoginition Procedure for Unconstrained Handwritten Words," *IWFHR-3, Buffalo New York, USA*, pp 122–131, May, 1993.

[3] P.G. De Luca, A. Gisotti, "Printed Character Preclassification Based on Word Structure," *Pattern Recognition*, Vol 24, No 7, pp 609–615, 1991.

[4] Sriganesh Madhavanath, Venu Govindaraju, "Holistic Lexicon Reduction," *IWFHR-3, Buffalo New York, USA*, pp 71–81, May, 1993.

[5] Thierry Paquet, Yves Lecourtier, "Recognition of Handwritten Sentences Using a Restricted Lexicon," *Patter Recognition*, Vol 26, No 3, pp 391–407, Mar, 1993.

[6] Charles Tuckey, "The Use of Global Recognition Techniques in Handwriting," *Technical Report, Department of Computer Science, University of Calgary, Canada*, Dec, 1993.