부분 매칭 방법을 이용한 효율적인 서식 문서 분류

변 영 철[†] · 최 영 우[††] · 김 경 환[†††] · 이 일 병[††††]

요 약

본 논문에서는 서식 문서를 짧은 처리 시간에 정확히 분류함으로써 실제 환경에서 응용할 수 있는 서식 분류 방법을 제안한다. 제안하는 방법은 서식 문서 이미지 전체를 다루기보다는 처리하고자 하는 서식 문서에서 서식 구조가 많이 다른 곳을 찾아서 매칭 영역으로 결정하고, 그 영역들에 대해서만 비교를 수행함으로서 계산 시간을 줄이고 인식률을 높인다. 선분 추출시 오류를 고려하기 위하여 기존 인쇄 문자와 채워진 데이터, 그리고 매칭 영역의 크기 정보를 페널티 함수로 반영하여 매칭 영역 선택시 고려한다. 본 방법은 구조적으로 많이 다르고, 양질의 특징을 포함하는 적은 수의 매칭 영역을 선택함으로써 처리 시간을 줄일 수 있음은 물론 높은 서식 분류율을 얻을 수 있다.

Efficient Form Document Classification Large using Partial Matching Method

Yung-Cheol Byun[†] · Yeong-Woo Choi[††] · Gyeong-Hwan Kim[†††] · Yill-Byung Lee[††††]

ABSTRACT

In this paper, we are proposing an efficient method of classifying form that is applicable in real life. Our method identifies a small number of matching areas by their distinctive images with respect to their layout structure and then by using a DP (Dynamic Programming) matching to match only these local regions. The penalty for each local area is computed by using the pre-printed text, filled-in data, and the size of the local area to prevent extracting erroneous lines. Our approach of searching and matching only a small number of structurally distinctive local regions would overcome the problems caused by the lengthy computation time and low recognition rate.

키워드 : Document Processing, Partial matching, Dp Matching, Disparity measure, Form Classification, Model-based approach

1. 서 론

1980년 이후 활발히 수행되고 있는 문서 처리에 관한 연구는 크게 두 갈래로 나뉜다. 하나는 문서에 존재하는 데이터를 데이터베이스에 자동으로 입력하고자 하는 데이터 자동 입력에 관한 연구이고, 또 다른 하나는 전자 문서 시스템을 위하여 종이 기반 문서를 컴퓨터 상의 문서로 자동으로 변환하는 연구이다. 전자는 사무 자동화의 관점에서 수행되었고, 후자는 전자 문서의 자동 생성이라는 관점에서 수행되었다.

본 연구는 전자에 관한 연구로서, 컴퓨터를 이용하여 처리하고자 하는 문서를 분류하고 문서에 있는 데이터를 인식한 후 데이터베이스에 자동으로 입력하기 위한 연구이다. 데이터를 인식하여 자동으로 입력하려면 서식 문서를 해석하는데 필요한 항목 이미지를 추출한 후 인식해야 한다. 처리하고자 하는 서식 문서의 유형이 여러 가지일 경우에는 한 가지 유형의 경우에 비하여 상대적으로 복잡한 알고리즘이 필요하며, 데이터를 인식하기에 앞서 서식 분류 과정이 필요하다.

본 논문은 지형적 구조를 가지는 서식 문서를 분류하는 방법을 제안한다. 특히 선분 등과 같은 그래픽 요소가 중요한 역할을 하는 문서를 분류하고자 한다. (그림 1-a)와 (그림 1-b)는 신용카드 매출표의 예로서, 기존 인쇄 문자(pre-printed text), 잡영, 그리고 채워진 데이터가 존재한다. 서식 문서 이미지에서 볼 수 있듯이, 선분과 항목 이미지, 그리고 기타 픽셀의 분포에 의해 서식 문서의 구조가 결정되며, 선분은 서식의 구조를 명확히 함으로써 서식을 쉽게 이해할 수 있도록 하는 역할을 한다.

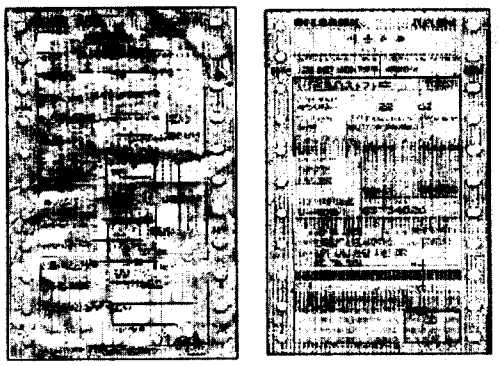
2. 관련 연구 및 고려 사항

2.1 관련 연구

서식 문서 분류와 관련된 주요 이슈로는 전처리, 특징 추

† 준 회 원 : 연세대학교 컴퓨터과학과 박사과정
†† 정 회 원 : 숙명여자대학교 정보과학부 교수
††† 정 회 원 : 서강대학교 전자공학과 교수
†††† 정 회 원 : 연세대학교 컴퓨터과학과 교수

논문접수 : 2000년 10월 31일, 심사완료 : 2000년 12월 18일



(a) (b)

(그림 1) 신용카드 매출표 서식 문서의 예

출, 그리고 인식기와 관련된 연구 등이 있다. 특징 추출은 특징 벡터의 차원을 줄임으로써 분류기의 부담을 줄일 수 있을 뿐만 아니라, 양질의 특징을 추출한다면 인식률과 처리 시간 면에서 분류기의 성능을 향상시킬 수 있다. 서식 문서를 분류하기 위해 사용하는 특징은 크게 물리적인 특징과 논리적인 특징으로 나눌 수 있다. 픽셀 매칭에서 사용할 수 있는 픽셀의 위치 정보, 서식 이미지의 크기와 픽셀 밀도 등은 전자의 예이다. 이에 반해 논리적인 정보로는 선분들의 상대적인 위치, 선분과 선분으로 구성되는 구조의 상대적인 위치, 서식 문서에 있는 항목들의 관계 정보 및 서식 문서 구성 요소들의 상대적인 위치 정보 등이 있을 수 있다.

[1]에서는 서식 인식을 위하여 선분으로 구성되는 10가지 특징점을 정의하여 이용하였다. 신경망 인식기를 사용하여 United States IRS(Internal Revenue Service) 서식 문서를 대상으로 실험한 결과 98%의 분류율을 얻었다. 실험 결과, 특징을 추출하는데 걸린 시간은 SPARCstation II를 이용하여 4.1 CPU초였다. [2]와 [3]에서는 수평 및 수직 선분 요소가 텍스트와 쉽게 구별되는 성질을 이용하였다. 구체적으로 선분의 위치와 두께 및 선분의 길이 정보를 추출하여 특징 벡터를 구성하였다. 이 연구에서는 DTW(Dynamic Time Wrapping) 알고리즘과 퍼지 매칭 방법을 이용하여 입력 서식 문서를 등록되어 있는 서식 문서의 하나로 분류하였다. [3]에서는 데이터베이스와 다섯 가지 모듈로 구성되는 모형 기반의 서식 처리 시스템을 제안하였다. 특징 벡터로서 수직선과 수평선 및 특징점을 이용하였으며, 모형 기반 매칭 방법을 이용한 서식 분류 알고리즘을 구현하였다. 실험 결과, 한 서식 문서를 모형화하는데 걸린 계산 시간은 대략 12초였다. [4]에서는 필드 항목의 중심 위치를 추출하여 비교하여 서식 문서를 분류하였다. [5]에서는 분류하고자 하는 서식 문서를 nm 영역으로 분할한 후, 각 영역에 대해 선분의 교차에 근거한 특징을 특징 벡터를 구성하여 서식 문서를 분류하였다.

[7]에서는 입력 서식 문서로부터 선분에 의한 9가지 특징점을 추출하여 문자열 표현으로 바꿈으로써 특징 벡터를 구성하였다. 인식기로서 간단한 거리 알고리즘을 이용하여 매칭을 수행하였다. 업무 서식 문서를 테스트 데이터로 사용하여 실험한 결과 한 서식을 처리하는데 1.75초의 CPU 초, 대략 25초 정도의 실제 처리 시간이 걸렸다. 서식 문서를 이해하기 위하여 연관 그래프(association graph)에 기반한 모형 매칭 방법이 제안되기도 하였다[8]. 이 연구에서는 선분과 관련된 정보를 특징 벡터로 추출하여 그래프로 표현함으로서 서식 분류 문제를 그래프 매칭 문제로 바꾸었다. 인식기로서 그래프 매칭 알고리즘을 사용하였으며, 14가지 유형의 서식 문서를 실험에 사용한 결과 평균 98%의 인식률을 얻었으며, 처리 시간은 5.76초였다.

일반적인 프로젝션 방법의 단점을 극복할 수 있는 방법으로서 스트립(strip) 프로젝션 방법이 제안되기도 하였다[10]. 이 연구에서는 분류 과정에 대해서는 설명하지 않았지만, 결과적으로 구해진 서식 구조는 서식 문서를 분류하는데 이용될 수 있다. [11]와 [12]에서는 선분에 의해 구성되는 모든 필드 항목을 인식한 후, 필드 항목의 좌상단 좌표를 이용하여 필드 항목들의 이웃 관계를 이진 트리를 이용하여 표현하였다. 이 트리는 입력 서식 문서를 분류하는데 이용될 수 있다.

[13]에서는 k-NN, MLP 및 트리 비교에 기반한 새로운 구조적 분류기를 이용하였다. 처리하고자 하는 서식 문서를 사각형 형태의 영역으로 분류한 후, 각 분할 영역의 데이터 픽셀의 밀도를 계산하였다. 계산된 밀도 정보는 k-NN와 MLP 인식기의 입력으로 사용하여 서식을 분류하였다. 또한 계층적 구조 정보를 추출한 후 트리로 표현하였으며, 이는 구조적 인식기의 입력으로 사용하였다. 세 가지 인식기를 사용한 결과 적절한 역치 값을 이용할 경우 87.31%에서 100%의 인식률을 얻을 수 있었다. 하지만 특징 추출을 위한 픽셀 기반 연산은 전반적으로 많은 처리 시간을 요구하였다[14, 15].

2.2 고려 사항

실제 환경에서 응용가능한 서식 처리 시스템을 개발하기 위해서는 몇 가지 해결해야 할 문제들이 있다. 서식 문서를 알려진 서식 중의 하나로 분류할 경우 서식 문서의 구조가 복잡하더라도 높은 인식률로 분류할 수 있어야 한다. 일반적으로 서식 문서 이미지의 크기는 개별 문자 이미지의 크기에 비해 훨씬 크며 서식 처리에 있어서 상대적으로 많은 계산 시간을 필요로 한다. 결과적으로, 시스템에 등록된 서식 문서의 종류가 증가할수록 서식 분석 및 처리에 필요한 시간이 증가한다. 또한 잡영과 이미지의 변형 및 왜곡에 대해서도 안정적으로 처리할 수 있어야 한다.

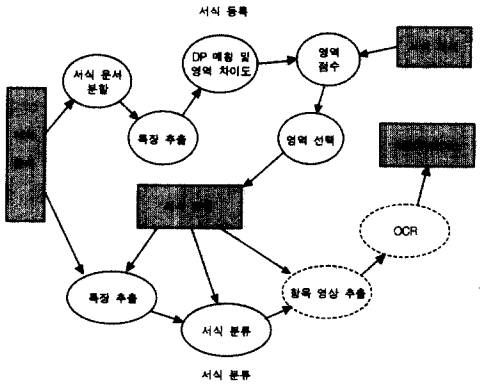
일반적으로 앞서 수행된 여러 연구 및 방법들은 처리하고자 하는 서식 문서 이미지의 모든 픽셀들을 동등하게 간

주하여 특징을 추출하였다. 일부 방법은 서식 구조를 인식하는데 좋은 결과를 보였으며, 그 결과 얻을 수 있는 서식 문서의 논리 구조는 서식을 해석하는데 유용하게 사용할 수 있었다. 그러나 서식의 구조가 크고 복잡한 구조를 갖는 문서인 경우 서식 처리 시간을 단축시킬 수 있는 방법이 필요하다. 이를 위해 본 연구에서는 서식 문서의 일부 영역에 대해서만 매칭을 수행하여 분류하는 방법을 제안하고자 한다. 이 방법은 입력된 서식 문서 영상에서 특징이 명확한 특정 영역에 대해서만 서식의 구조를 인식한 후, 서식 분류를 수행함으로써 처리 시간을 대폭 줄일 수 있고, 또한 인식률을 향상시킬 수 있다.

3. 제안하는 서식 분류 방법

3.1 모형 기반 방법

본 논문에서 짧은 시간에 높은 인식률로 서식 문서를 분류하기 위하여 (그림 2)와 같은 모형 기반 방법을 제안한다. 서식을 분류하기 위한 과정은 서식 등록 과정과 서식 분류 과정으로 구성된다. 서식 등록 단계에서는 처리하고자 하는 서식 문서에서 구조적인 특징 차이가 많은 영역을 찾아 매칭 영역으로 등록한다. 서식 분류 단계에서는 앞에서 찾은 매칭 영역에서 DP 매칭을 이용하여 영역을 비교한다. 본 논문에서는 (그림 2)에서 점선으로 표시된 항목 영상 추출과 OCR을 제외한 나머지 부분에 대해 설명한다.



(그림 2) 제안하는 모형 기반 서식 처리 시스템

서식 처리 과정을 요약하면 다음과 같다. 우선, 등록하고자 하는 모든 서식 문서에서 선분을 추출한 후 선분의 분포를 고려하여 문서를 분할한다. 문서를 분할한 후 가능한 서식 문서 쌍에 대해 서로 대응되는 위치의 분할 영역의 구조적인 차이를 의미하는 영역 차이도(disparity)를 정의하여 계산한다. 이 경우 영역 차이도는 매칭 영역에 존재하는 선분을 특징으로 DP 매칭을 수행하여 구한다. 잡영 및 텍스트 등에 의해 선분이 잘못 추출될 수도 있기 때문에 영역 차이도 이외에 기존 인쇄 문자와 채워진 데이터, 그리고 영역 크기에 관한 지식을 페널티 함수로 표현하여 매칭 영역 선택시 반영한다. 결국, 매칭 영역 선택시 우선 순위를 결정하기 위한 영역 점수(score)는 영역 차이도 값과 페널티의 함수로 결정된다. 최종적으로 영역 점수에 근거하여 선택된 매칭 영역에 대한 정보인 매칭 영역의 위치 및 매칭 영역에 있는 선분에 의한 서식 특징 정보가 서식의 모형으로 등록되며, 이를 이용하여 서식 처리 과정에서 입력된 서식 문서를 분류한다.

본 연구에서 달성하고자 하는 목표는 다음과 같다.

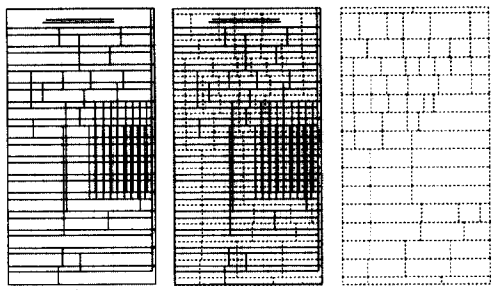
- 서식 문서를 분류함에 있어서 서식 문서 이미지 전체가 아니라 구조적으로 많이 다른 일부 영역 만을 추출하여 비교함으로써 처리 시간을 단축하고 인식률을 높인다.
- 서식 구조가 복잡하고 유사한 문서에 대하여 구조적으로 많이 다른 부분을 위주로 비교함으로써 효과적으로 분류할 수 있도록 한다.
- 기존 인쇄 문자와 채워진 데이터 뿐만 아니라 잡영이 포함되는 서식 문서 이미지를 효율적으로 분류함으로써 실제 환경에서 적용할 수 있는 서식 처리 시스템을 개발할 수 있도록 한다.

4. 특징 추출

서식 문서의 구조를 분석한 후 선분의 분포를 고려하여 문서를 분할한다. 문서 분할은 서식 문서의 특징을 안정적으로 추출할 수 있도록 수행해야 한다. 그리고 분할된 영역들은 적어도 하나의 선분을 포함함으로써 비교가 가능하도록 한다. 또한, 허용된 범위 내에서 서식 문서가 이동된 경우에도 안정적으로 특징을 추출할 수 있도록 분할해야 한다. 이와 같은 사항을 만족하도록 하기 위하여 서식 문서에 있는 선분의 위치 및 시작점/끝점의 위치를 고려한다. 수직 선분과 수평 선분 각각에 대하여, 우선 서로 인접하는 선분 중 가장 멀리 떨어져 있는 두 선분의 중간 부분을 분할한다. 다음, 그 결과 생성되는 두 블록 각각에 대해 동일한 과정을 반복한다. 이러한 서식 분할 과정은 인접하는 두 선분의 거리가 특정한 값(분할 상수)보다 작을 때까지 계속된다.

(그림 3)은 서식 문서를 분할한 결과이다. (그림 3-a)는 처리하고자 하는 두 개의 서식 문서를 겹친 모습이고, (b)의 점선은 특정 분할 상수를 이용하여 서식 문서를 분할한 결과이다. 결과적으로 (c)와 같은 최종 분할 결과를 얻을 수 있다.

서식 문서에서 항목의 위치와 분포는 서식을 해석하는데 있어 중요하다. 선분은 항목들을 쉽게 찾고 해석할 수 있도록 하며 서식을 구분하는 중요한 단서가 된다. 따라서 선분 정보를 이용하여 특징 벡터를 구성한다. 특징 벡터는



(그림 3) 서식 문서를 분할한 예

수평/수직 선분의 위치 및 시작점/끝점으로 구성된다. 예를 들어 하나의 수직 선분은(v_l, v_s, v_e)으로 표현된다. v_l은 수직 선분의 위치인 x 좌표를 의미하며, v_s와 v_e는 각각 y 좌표로 표현되는 수직 선분의 시작점과 끝점을 의미한다. 따라서 하나의 분할된 영역에 m개의 수직 선분과 n개의 수평 선분이 존재할 경우 특징 벡터는 다음과 같이 구성된다.

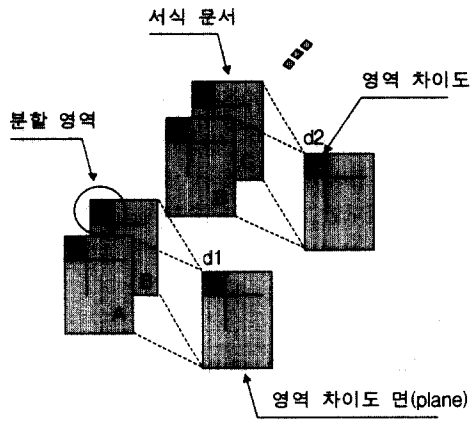
$$((v_{l1}, v_{s1}, v_{e1}),(v_{l2}, v_{s2}, v_{e2}), \mathrm{K},(v_{lm}, v_{sm}, v_{em}))$$
$$((h_{l1}, h_{s1}, h_{e1}),(h_{l2}, h_{s2}, h_{e2}), \mathrm{K},(h_{lm}, h_{sm}, h_{em}))$$

5. 매칭 영역 결정 및 서식 분류

5.1 DP 매칭에 의한 영역 차이도 값 계산

본 절에서는 앞서 분할한 각 영역에서 추출한 특징을 이용하여 영역 차이도 값을 구하고, 이를 이용하여 매칭 영역을 선택하는 방법에 대해 설명한다. 매칭 영역을 선택할 경우, 해당 매칭 영역에 대해서만 비교를 수행하더라도 적절한 시간 내에 입력 서식 문서를 유일하게 분류할 수 있어야 한다. 따라서 지형적 구조의 차이가 많은 영역을 위주로 가급적 적은 수의 매칭 영역을 선택할 필요가 있다. 본 연구에서는 서식 문서 분류시 선분 정보를 이용할 것이므로 다음 사항을 고려해야 한다. (1) 선분과 유사한 잡영이 있을 수 있다. (2) 선분 전체가 사라질 수 있다. (3) 하나의 선분이 두 개 이상의 선분으로 분할 될 수 있다. (4) 선분의 일부가 사라질 수 있다. 본 연구에서는 이와 같은 문제점들을 해결하기 위해서 DP 매칭 방법을 이용하여 영역 차이도 값을 계산한다.

영역 차이도 값은 (그림 4)와 같이 두 서식 문서 간 분할된 영역에서의 구조적인 특징의 차이를 거리로 정의한 값이다. 서식 문서를 n개의 영역으로 분할한다고 가정하자. 처리하고자 하는 서식 문서 유형이 두 개일 경우 n개의 영역 차이도 값이 계산된다. 처리하고자 하는 서식 문서가 세 개일 경우 세 쌍에 대해 n개의 영역 차이도 값을 구할 수 있다. 이 때 각 쌍에 대해 구한 n개의 영역 차이도 값을 일컬어 (그림 4)에서 처럼 영역 차이도 면(disparity plane)이라고 부르기로 한다. 영역 차이도 벡터는 모든 영역 차이도 면에서 서로 대응되는 위치에 있는 영역 차이도 값을 구성요소로 벡터를 구성한 것이다.



영역 차이도 벡터 d = (d1, d2, ···, dn)

(그림 4) 영역 분할 및 영역 차이도 값

두 개의 특징 벡터의 거리인 영역 차이도 값은 식 (1)의 DP 매칭 알고리즘을 이용하여 계산한다.

$$g(i,j) = \min\begin{Bmatrix} g(i-1,j)+C \\ g(i-1,j-1)+d(i,j) \\ g(i,j-1)+C \end{Bmatrix} \quad (1)$$

이 경우, i와 j는 벡터의 인덱스를 의미하며, C는 DP 매칭 상수이다. DP 매칭 공식의 g(i, j)에 의해 가중치 그래프의 가중치 w_{ij}가 구해진다. 비교되는 두 특징 벡터의 요소의 수가 각각 m, n인 경우 $1 \le i \le m$, $1 \le j \le n$을 만족한다. d(i, j)는 두 선분 사이의 거리를 의미한다. 예를 들어, 처리하고자 하는 서식 문서의 유형이 두 개이고, 각 매칭 영역에 대한 특징 벡터가 각각 ((a_{l1}, a_{s1}, a_{e1}), (a_{l2}, a_{s2}, a_{e2}), ···, (a_{lm}, a_{sm}, a_{em}))와 ((b_{l1}, b_{s1}, b_{e1}), (b_{l2}, b_{s2}, b_{e2}), ···, (b_{ln}, b_{sn}, b_{en}))인 경우 i번째 선분과 j번째 선분의 거리 d(i, j)는 식 (2)와 같이 선분의 위치와 시작점 및 끝점의 위치, 그리고 길이를 고려하여 계산한다.

$$d(i,j) = |a_{li} - b_{lj}| + \alpha(|a_{si} - b_{s2}| + |a_{ei} - b_{e2}|) \quad (2)$$

식 (2)에서 α는 시작점/끝점에 의한 거리가 수식에 어느 정도 반영되는지를 의미하는 비례 상수이다. 선분 정보의 DP 매칭 알고리즘을 이용함으로써 서로 매칭되는 두 선분의 거리가 영역 차이도 값에 더해진다. DP 매칭을 수행한 결과 가중치 그래프가 생성되고, 가중치 그래프에서 식 (3)을 만족하는 최단 경로, k_1, k_2, ···, k_Q를 찾아 경로에 있는 모든 가중치를 더한 결과가 영역 차이도 값이 된다.

$$disparity = \min\left(\sum_{i=0}^{Q} w(i)\right) \quad (3)$$

식 (4)에서 w(i)는 k_i 노드의 가중치를 반환하는 함수이다. 따라서 영역 차이도는 (0,0) 노드와 (m, n) 노드를 연결하는 최소 경로 상에 있는 가중치를 합한 결과이다. 서식 문서의 수가 n개일 경우 ${}_nC_2$ 개의 문서 쌍이 존재할 수 있으며, 각각의 분할된 영역에 대해 아래와 같은 영역 차이도 벡터를 구할 수 있다. 분할된 영역의 수가 m개인 경우 m개의 영역 차이도 벡터를 얻을 수 있다.

$$d=(d_1, d_2, \cdots, d_{nC_2})$$

5.2 영역 점수 계산 및 매칭 영역 선택

5.2.1 영역 차이도 값의 정규화

서식 분류시 비교할 매칭 영역을 선택하기 위하여 영역 차이도 벡터를 이용하여 각 영역 별로 점수를 계산한다. 매칭 영역 선택에 있어서 중요한 기준은 서식 분류율이므로, 서식 문서의 구조적인 차이를 의미하는 영역 차이도 값을 이용하여 서식 문서들을 쉽게 구분할 수 있는 영역을 위주로 선택한다. <표 1>은 분할된 4개의 영역에서 각 매출표 간의 구조적 차이인 영역 차이도 값을 구한 예이다. <표 1>의 경우 A 문서와 B 문서를 가장 잘 분류할 수 있는 영역은 a_2 영역이다. 또한, a_1 영역은 B 문서와 C 문서를 가장 잘 분류할 수 있는 영역이다. 서식 문서 C와 A의 경우 a_4 영역이 구조적 차이가 가장 많이 나는 영역이다.

〈표 1〉 분할된 영역에 대한 영역 차이도 값의 예

area	AB	BC	CA
a_1	0.10	0.50	0
a_2	0.20	0.40	0.10
a_3	0.10	0	0.70
a_4	0	0.10	0.80

다시 말해, AB를 가장 잘 분류할 수 있는 영역은 a_2이고, CA를 가장 잘 분류할 수 있는 영역은 a_4이다. 여기에서 a_2 영역의 영역 차이도 값 0.20은 a_4영역의 영역 차이도 값 0.80에 비해 상대적으로 작아도 모든 서식 문서를 분류하기 위해서는 동등하게 고려해야 한다. 따라서, 두 서식 문서를 가장 잘 분류하는 영역의 영역 차이도 값이 1이 되도록 각각의 벡터 요소를 정규화한다. <표 1>의 각 열에 대해 식 (4)를 이용하여 정규화를 수행한다.

$$d_{N_{ij}}=\frac{d_{ij}}{d_{Mj}} \qquad 1\le i\le m,\ 1\le j\le {}_nC_2 \tag{4}$$

이 경우 d_{ij}는 i번째 영역에 대한 영역 차이도 벡터의 j번째 벡터 요소를 의미하며, d_{Mj}는 모든 영역에 대한 j번째 벡터 요소 중 최대 영역 차이도 값을 표현한다. m과 n은 영역의 수와 처리하고자 하는 서식 문서의 수를 의미한다. 결과적으로, 두 서식 문서를 가장 잘 분류할 수 있는 분할 영역의 영역 차이도 값은 1로 정규화가 되며, 전혀 구분할 수 없는, 혹은 서식 구조가 동일한 영역의 영역 차이도 값은 0이 된다. <표 2>는 영역 차이도를 정규화한 결과이다.

〈표 2〉 영역 차이도 값을 정규화한 결과

area	AB	BC	CA
a_1	0.50	1.00	0
a_2	1.00	0.80	0.13
a_3	0.50	0	0.88
a_4	0	0.20	1.00

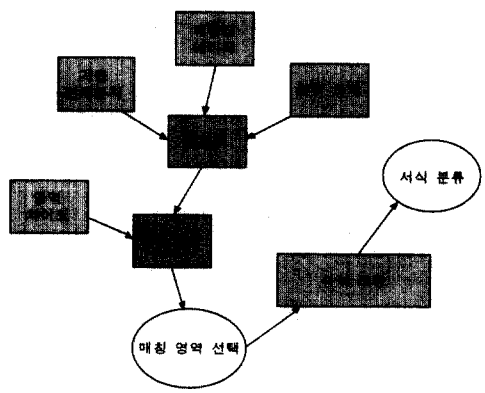
5.2.2 지식을 이용한 영역 점수 계산

분할한 영역 중 서식 문서를 잘 분류할 수 있는 매칭 영역을 선택하기 위하여 각 분할 영역에 영역 점수를 계산함으로써 순서를 부여한다. 가장 간단한 방법은 각 영역 차이도 벡터의 영역 차이도 값을 모두 더한 후 평균값을 영역 점수로 이용하여 상대적으로 큰 점수를 갖는 영역을 매칭 영역으로 선택하는 것이다. 식 (5)는 평균값을 이용한 영역 점수인 $S_{avg(i)}$를 구하는 공식이다.

$$S_{avg(i)}=\frac{1}{{}_nC_2}\sum_{j=1}^{{}_nC_2} d_{Nij} \tag{5}$$

위의 식에서 $S_{avg(i)}$는 i번째 매칭 영역의 점수를 의미하며, $0 \le S_{avg(i)} \le 1$을 만족한다.

영역 차이도 값은 선분의 위치 및 길이 정보를 이용하여 구한 결과이므로 식 (5)에 의해 계산된 영역 점수는 선분만을 고려한 결과이다. $S_{avg(i)}$ 값이 큰 순서대로 매칭 영역을 선택할 경우 매칭 영역에 기존 인쇄 문자와 채워진 데이터 등이 존재하여 잘못된 선분이 추출되면 서식 분류시 오인식의 원인이 될 수 있다. 따라서 기존 인쇄 문자와 채워진 데이터가 있는 영역을 가급적 회피하도록 매칭 영역을 선택할 필요가 있다.



(그림 6) 패널티를 이용한 영역 점수 조정

따라서 (그림 6)과 같이 선분의 추출에 영향을 미칠 수

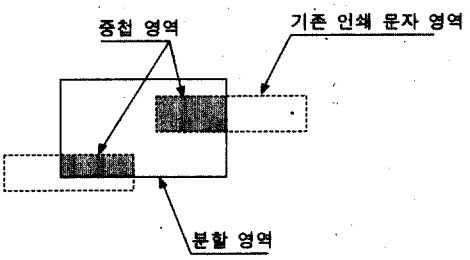
있는 요소인 매칭 영역의 크기, 해당 매칭 영역에 기존 인쇄 문자와 채워진 데이터가 어느 정도 존재하는지를 패널티로 구한 후, 이를 이용하여 앞서 구한 $s_{avg(i)}$ 값을 재조정한다. 이를 위한 공식은 식 (6)과 같다.

$$s_i = d_{avg(i)} - p_i$$
$$p_i = \beta_2 r_{pi} + \beta_2 r_{ti} + \beta_3 \left(1 - \frac{a_i}{a_M}\right) \quad (6)$$

식 (6)에서 r_{pi}는 i번째 영역 내에 기존 인쇄 문자가 어느 정도 존재하는지를 의미하는 중첩 비율 값이며, r_{ti}는 채워진 데이터가 어느 정도 존재하는지를 의미하는 값이다. 그리고 a_i와 a_M은 각각 i 번째 영역의 면적과 모든 영역의 면적 중에서 최대의 면적을 의미한다. β_1, β_2, 그리고 β_3은 각 항을 어느 정도 반영할지를 의미하는 비례 상수이다.

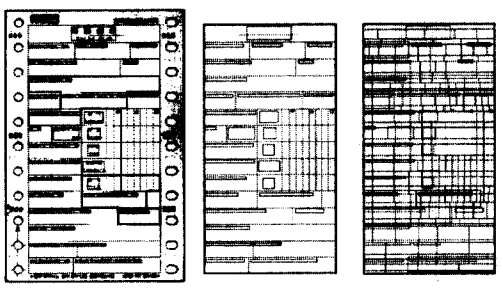
분할된 영역에 기존 인쇄 문자가 존재할 경우 중첩 비율을 계산하는 공식은 식 (7)과 같다. 이 경우 n은 처리하고자 하는 서식 문서의 수를 의미하며, a_o는 분할된 영역과 기존 인쇄 문자 영역의 중첩된 면적을 의미한다. a_I은 (그림 7)과 같이 분할된 영역의 면적을 의미한다.

$$r = \frac{1}{N} \times \frac{a_o}{a_I} \quad (7)$$



(그림 7) 분할 영역과 기존 인쇄 문자 영역의 중첩

(그림 8-a)는 비어있는 이지체크 신용카드 매출표에서 기존 인쇄 문자를 추출한 모습이다. 결과적으로 얻을 수 있는 기존 인쇄 문자 지식은 (b)와 같으며, (c)는 분할 영역과 기존 인쇄 문자 영역을 겹친 모습을 나타낸다.



(그림 8) 기존 인쇄 문자 영역과 분할 영역

마찬가지로 채워진 데이터의 경우에도 동일한 방법으로 중첩 비율 값을 구할 수 있다. 단, 기존 인쇄 문자 영역은 비어있는 매출표의 서식 구조를 인식한 후 사각형 형태의 셀 영역 내부를 검사함으로써 자동적으로 찾을 수 있지만, 채워진 데이터의 경우에는 빈 서식과 채워진 서식을 비교하여 차이가 있는 셀 영역을 채워진 데이터 영역으로 구한다. 위의 수식에 의하여 영역에 기존 인쇄 문자와 채워진 데이터가 많이 존재하고 매칭 영역의 크기가 작을수록 패널티 값은 커진다. 결과적으로 i번째 매칭 영역의 영역 점수 값인 s_i는 상대적으로 작아져서 매칭 영역으로 선택될 가능성이 줄어든다.

5.2.3 영역 점수를 이용한 매칭 영역 선택

마지막으로, 앞 절에서 계산한 영역 점수 값을 이용하여 서식 분류시 사용할 매칭 영역을 선택한다. 분할된 영역에 대한 영역 차이도와 패널티, 그리고 영역 점수 값이 <표 3>과 같을 때, 첫번째로 선택되는 칭 영역은 a_2이다. 선택할 매칭 영역의 수가 두 개일 경우 a_2와 a_1이 선택된다. 따라서 매칭 영역이 선택되는 순서는 a_2, a_1, a_3, a_4이다.

〈표 3〉 패널티를 이용한 영역 점수 조정 결과

area	AB	BC	CA	s_{avgi}	p_i	s_i
a_1	0.50	1.00	0	0.50	0.05	0.45
a_2	1.00	0.80	0.13	0.64	0.10	0.54
a_3	0.50	0	0.88	0.46	0.12	0.34
a_4	0	0.20	1.00	0.40	0.10	0.30

6. 실험 결과

6.1 실험 환경 및 내용

본 연구에서 제안한 방법을 테스트하기 위하여 여섯 가지 유형의 신용카드 매출표 서식을 실험에 사용하였다. 신용카드 매출표 서식은 동일한 목적으로 만들어진 문서이고, 문서의 크기와 구조가 서로 비슷함으로써 제안한 방법을 검증하는데 적합하기 때문이다. <표 4>와 같이, 여섯 가지 서식 문서에 대해 모두 126개의 서식 이미지를 200dpi 해상도로 스캔받아 실험에 사용하였으며, 이중 여섯개의 빈 매출표 이미지는 모형을 등록하기 위해 사용하였고, 120개의 이미지는 테스트에 사용하였다. 이 때 매출표 이미지의 평균 크기는 8211146 픽셀이었다. 서식 분류 시스템은 Pentium PC (PII 366)에서 C++ 언어를 이용하여 구현하였다.

〈표 4〉 실험 데이터

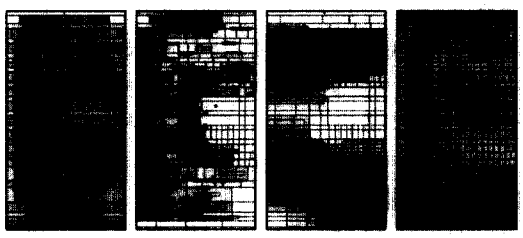
서 식 문 서	학 습	테 스 트
A 사	1	20
B 사	1	20
C 사	1	20
D 사	1	20
E 사	1	20
F 사	1	20
합 계	6	120

서식 등록 단계에서는 여섯 가지 유형의 채워지지 않은 신용카드 매출표를 이용하여 서식 문서의 구조를 인식한다. 그 다음, 인식한 선분의 분포를 고려하여 6개의 서식 문서를 동일하게 분할한다. 분할한 각 영역에 대해 DP 매칭 방법을 이용하여 서식 문서 사이의 구조적 거리를 의미하는 영역 차이도 값을 구한 후, 이를 이용하여 매칭 영역을 선택한다. 각각의 신용카드 매출표에 대해 매칭 영역을 선택한 후 영역에 있는 선분에 의한 구조적 정보를 모형으로 등록한다.

서식 분류 단계에서는 사용자에 의해 채워진 120개의 신용카드 매출표 이미지에 대해 등록된 매칭 영역에 대해서만 특징을 추출한 후, 모형으로 등록되어 있는 특징과 DP 매칭을 수행하여 등록되어 있는 모형 중 하나로 분류한다. 또한 매칭 영역의 수에 따라 서식을 분류하는데 필요한 처리 시간을 알아본다.

6.2 서식 분할 및 영역 차이도 값 계산

서식 문서 분할 단계에서는 수평 분할선과 수직 분할선을 찾는다. 이를 위해, 분할 상수 d_h와 d_v를 정의한다. d_h는 수직 분할선을 구하기 위한 분할 상수이다. 가령, d_h가 7일 경우, 수직 선분의 위치(x 좌표)와 수평 선분의 시작점과 끝점의 위치(x 좌표) 중 서로 이웃하는 거리가 7 이상일 경우 중간 부분을 분할한다. 마찬가지로, d_h는 수평 분할선을 구하기 위한 상수로서, 수평 선분의 위치(y 좌표)와 수직 선분 시작점과 끝점의 위치(y 좌표) 중 서로 이웃하는 좌표의 거리가 d_h보다 클 경우 중간 부분을 분할한다. (그림 9)는 분할 상수 d_h/d_v가 13/13일 경우 분할한 결과 및 해당 영역에서의 영역 차이도 값의 평균값인 $s_{avg(i)}$를 표시한다(a). 그리고 해당 영역에서의 기존 인쇄 문자의 중첩 비율 및 채워진 데이터 중첩 비율(c)과 영역의 크기를 표시한 것이다. (a)의 경우, 흰 색의 분할 영역은 구조적으로 차이가 없는 곳으로 영역 점수가 0에 가까우며, 이와는 반대로, 검정 색의 영역은 영역 점수가 1에 가까우며 구조적으로 차이가 큰 영역을 의미한다. 두 서식 문서 간 구조적 정보의 차가 크면 클수록 영역 차이도 값도 커진다. 본 실험에서 분할 상수 13/13은 실험에 근거하여 인식률 면에서 가장 적합한 값을 선택하였으며, 이 경우 분할된 매칭 영역의 수는 276 개였다.



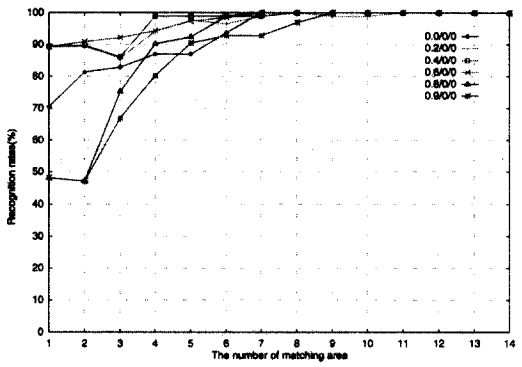
(a) (b) (c) (d)

(그림 9) 영역 차이도 값과 지식에 대한 정보 :
(a)영역 차이도 값, (b)기존 인쇄 문자,
(c)채워진 데이터의 중첩 비율,
(d)영역의 크기

6.3 영역 점수를 이용한 영역 선택 및 서식 분류

분할된 영역에 대해 영역 점수를 계산한 후, 영역 점수의 크기 순서로 매칭 영역을 선택하여 서식을 분류하였다. 이 경우 페널티를 구성하는 기존 인쇄 문자, 채워진 데이터, 그리고 영역의 크기에 대한 비례 상수 β_1, β_2, β_3을 다르게 설정하여 실험함으로써 (1) 기존 인쇄 문자가 서식 분류율에 미치는 영향, (2) 채워진 데이터가 인식률에 미치는 영향, 그리고 (3) 영역의 크기가 인식률에 미치는 영향을 분석하였다.

먼저, 패널티를 구하는 공식에서 비례 상수 β_1, β_2, β_3이 0으로 설정하여 영역 점수를 계산하였다. 계산된 영역 점수 값의 크기 순서로 14개의 매칭 영역을 선택하여 신용카드 매출표 이미지를 분류하였다. 매칭 영역의 수를 14개로 제한한 이유는 실험에 근거하여 100%의 분류율을 얻는데 충분하다고 판단하였기 때문이다. 기존 인쇄 문자가 인식률에 미치는 영향을 알아보기 위하여 기존 인쇄 문자의 비례 상수인 β_1을 증가하면서 실험한 결과 (그림 10)과 같은 결과를 얻었다. (그림 10)에서 0.0/0/0은 β_1, β_2, β_3이 각각 0, 0, 0인 경우를 의미하며, 이는 기존 인쇄 문자, 채워진 데이터, 영역 크기를 전혀 반영하지 않은 경우이다.

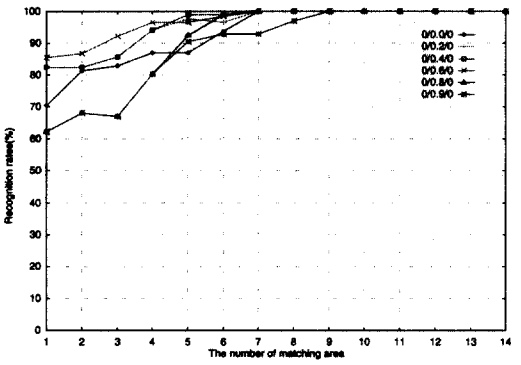


(그림 10) 기존 인쇄 문자의 반영 정도에 따른 서식 분류율

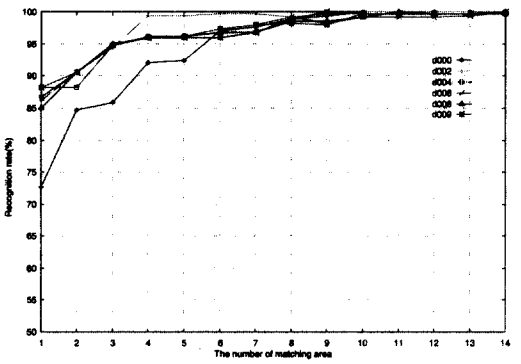
실험 결과, 비례 상수 β_1이 0.2, 0.4, 0.6일 경우 분류율이 향상됨을 알 수 있었다. 기존 인쇄 문자가 있는 영역을 회피할 경우 그렇지 않은 경우(0.0/0/0)에 비해 분류율이 향상되었다. 하지만 β_1이 0.8, 0.9일 경우에는 오히려 인식률이 낮아졌다. 왜냐하면, score(s_i)를 계산할 경우 β_1값이 크면 영역 차이도 값에 의한 $s_{avg(i)}$가 상대적으로 작게 반영되어 선분에 의한 구조적인 차이가 많은 부분이 매칭 영역으로 선택되지 않기 때문이다.

(그림 11)은 채워진 데이터가 있는 영역이 매칭 영역에서 제외되도록 선택한 후 분류 실험을 한 결과이다. 실험 결과

채워진 데이터가 상대적으로 없는 영역을 매칭 영역으로 선택할 경우 서식 분류율이 향상됨을 알 수 있었다. 하지만, 이 결과는 기존 인쇄 문자를 고려할 경우와 마찬가지로, 비례 상수가 특정 값보다 크면 상대적으로 영역 차이도 반영 비율이 감소하여 인식률이 오히려 떨어짐을 알 수 있었다.



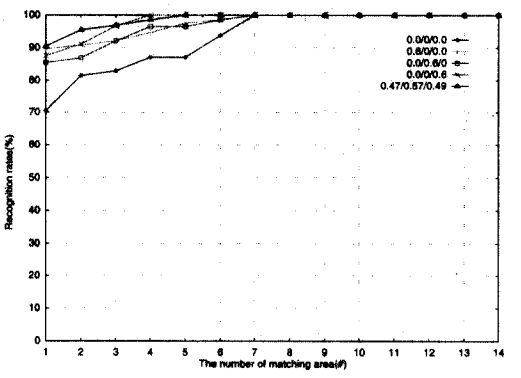
(그림 11) 채워진 데이터의 반영 정도에 따른 서식 분류율



(그림 12) 영역의 크기에 따른 서식 분류율

(그림 12)는 매칭 영역의 크기가 인식률에 어떤 영향을 미치는지를 알아보기 위하여 영역 크기 비례 상수인 β_3을 증가하면서 분류 실험을 수행한 결과이다. 매칭 영역 선택시 영역의 크기를 고려한 결과가 고려하지 않은 결과(0.0/0/0)에 비해 인식률이 전반적으로 향상되었다. 특히 비례 상수 β_2가 0.2일 때 좋은 결과를 얻을 수 있었다. 그리고 비례 상수가 0.2보다 클 경우에는 인식률에 큰 변화가 없었다. 이는 영역의 크기가 어느 정도 이상일 경우에는 본 실험에서 사용한 선분 추출 알고리즘이 선분을 안정적으로 추출한다는 것을 의미한다.

마지막으로, 매칭 영역 선택시 선분 추출에 영향을 줄 수 있는 세 가지 요소를 모두 고려하여 실험하였으며, 성능 비교를 위하여 앞서 구한 결과와 비교하였다. (그림 13)은 영역 차이도 값이 크면서 상대적으로 기존 인쇄 문자와 채워진 데이터가 없고 영역의 크기가 큰 영역을 매칭 영역으로 선택하여 실험한 결과이다. 실험 결과 β_1, β_2, β_3이 각각 0.47, 0.57, 0.49일 경우 좋은 결과를 얻을 수 있었다. 결과적으로 지식에 의해 영역 점수를 재조정하여 매칭 영역을 선택함으로써 인식률을 향상시킬 수 있었다.



(그림 13) 모든 지식을 반영할 경우의 서식 분류율

7. 결 론

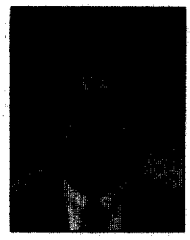
본 논문에서는 지식을 사용한 부분 매칭 방법을 이용하여 다양한 유형의 서식 문서를 효율적으로 처리할 수 있는 새로운 방법을 제안하였다. 제안한 방법은 영역 점수에 근거하여 매칭 영역을 선택한다. 중복되는 영역과 잡영 및 채워진 데이터가 있는 영역을 선택하지 않고 양질의 특징을 포함하는 매칭 영역을 선택한다. 처리하고자 하는 서식 문서들이 구조적인 차이가 큰 영역을 위주로 가급적 적은 수의 매칭 영역을 선택함으로써 빠른 시간 내에 높은 인식률로 서식을 처리할 수 있다.

실험 결과 제안하는 방법은 처리해야 할 서식 문서의 구조가 서로 유사하고, 서식 문서에 잡영 및 채워진 데이터가 있는 경우에도 효과적으로 적용될 수 있음을 알 수 있었다. 비교적 서식 구조가 유사한 신용카드 매출표 여섯 가지 문서를 이용하여 실험한 결과 적은 수의 매칭 영역만으로도 높은 인식률로 서식 문서를 분류할 수 있었다. 처리하고자 하는 서식 문서에 대한 지식과 영역 차이도 값을 이용하여 양질의 특징을 포함하는 영역을 위주로 비교를 수행함으로써 적절한 시간 내에 높은 인식률로 서식을 처리할 수 있었다. 영역 점수를 이용함으로써 단지 4~5개의 매칭 영역을 비교하더라도 100%의 인식률을 얻을 수 있었으며, 이 경우 1개의 신용카드 매출표를 처리하는데 0.7~0.8초 정도의 계산 시간이 필요하였다. 이는 실제 환경에서 적용가능한 시스템 개발에 있어서 인식률 및 처리 시간 관점에서 고무적인 결과로 여겨진다.

참 고 문 헌

[1] S. L. Taylor and R. Fritzson, J. A. Pastor, "Extraction of data from preprinted forms," International Workshop on

Machine Vision Applications, Vol.5, pp.211-222, 1992.
[2] R. G. Casey, D. R. Ferguson, K. Mohiuddin and E. Walach, "Intelligent forms processing system," International Workshop on Machine Vision Applications, Vol.5, pp.511-529, 1992.
[3] J. Mao, M. Abayan and K. Mohiuddin, "A Model- Based Form Processing Sub-System," International Conference on Document Analysis and Recognition, pp.691-695, 1996.
[4] S. Shimotsuji and M. Asano, "Form Identification based on Cell Structure," International Conference on Document Analysis and Recognition, pp.793-797, 1996.
[5] S. W. Lam, L. Javanbakht and S. N. Srihari, "Anatomy of a form reader," International Conference on Document Analysis and Recognition, pp.506-509, 1993.
[6] T. Watanabe, Q. Luo and N. Sugie, "Layout Recognition of Multi-Kinds of Table-form Documents," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.17, No.4, pp.432-445, 1995.
[7] A. Ting, M. K. Leung, S.-C. H and K.-Y. Chan, "A Syntactic Business Form Classifier," International Conference on Document Analysis and Recognition, pp.301-304, 1995.
[8] Y. Ishitani, "Model Matching Based on Association Graph for Form Image Understanding," International Conference on Document Analysis and Recognition, pp.287-292, 1995.
[9] Y. Hirayama, "A Method for Table Structure Analysis using DP Matching," International Conference on Document Analysis and Recognition, pp.583-586, 1995.
[10] Jiun-Lin Chen and Hsi-Jian Lee, "A Novel Form Structure Extraction Method Using Strip Projection," International Conference on Document Analysis and Recognition, pp. 823-827, 1996.
[11] T. Sobue and T. "Watanabe, Identification of Item Fields in Table-form Documents with/without Line Segments," International Workshop on Machine Vision Applications, pp.522-525, 1996.
[12] T. Watanabe, "Document Analysis and Recogntion," IEICE Trans. Inf. & Syst., Vol.E82-D, No.3, pp. 601-610, 1999.
[13] P. Heroux, S. Diana. A. Ribert and E. Trupin, "Classification Method Study for Automatic Form Class Identification," International Workshop on Frontiers in Handwriting Recognition, pp.926-928, 1998.
[14] Y. Hirayama, "Analyzing Form Images by Using Line-Shared-Adjacent Cell Relations," International Conference on Document Analysis and Recognition, pp.768-772, 1996.
[15] T. Sobue and T. Watanabe, "Identification of Item Fields in Table-form Documents with/without Line Segments," International Workshop on Machine Vision Applications, pp.522-525, 1996.



변 영 철

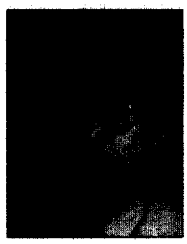
e-mail : bcart@csai.yonsei.ac.kr
1993년 제주대학교 정보공학과 졸업(학사)
1995년 연세대학교 대학원 컴퓨터과학과 졸업(석사)
1997년 연세대학교 대학원 컴퓨터과학 산업시스템 공학과 박사과정 수료
1997년~현재 연세대학교 대학원 전문연구원
관심분야 : 문서인식, 패턴인식, 신경회로망, 데이터마이닝 등

최 영 우

e-mail : ywchoi@sookmyung.ac.kr
1985년 연세대학교 전자공학과 졸업(학사)
1986년 University of Southern California 컴퓨터공학과 졸업(석사)
1994년 University of Southern California 컴퓨터공학과 졸업(박사)
1994년~1997년 LG전자기술원 선임연구원
1997년~현재 숙명여자대학교 정보과학부 조교수
관심분야 : 영상처리, 패턴인식, 문자인식 등

김 경 환

e-mail : gkim@ccs.sogang.ac.kr
1984년 서강대학교 전자공학과 졸업(학사)
1986년 서강대학교 대학원 전자공학과 (공학석사)
1996년 미국 SUNY at Buffalo 전기 및 컴퓨터 공학과(공학박사)
1986년~1991년 금성전기(정밀) 기술연구소 선임연구원
1993년~1997년 CEDAR/SUNY at Buffalo Research Scientist
1997년~현재 서강대학교 전자공학과 조교수
관심분야 : 영상신호해석, 패턴인식, 신경회로망, Embedded System Design



이 일 병

e-mail : yblee@csai.yonsei.ac.kr
1976년 연세대학교 전자공학과 졸업(학사)
1980년 University of Illinois 전산과학과 졸업(석사)
1985년 University of Massachusetts 전산정보과학과 졸업(박사)
1985년~현재 연세대학교 컴퓨터과학과 교수
1995년~1999년 연세대학교 소프트웨어 응용 연구소 소장
관심분야 : 신경회로망, 문서인식, Computer Vision, Data Mining