부분 매칭을 이용한 서식 이해에 관한 연구*

변영철○ 윤성수 김경환† 최영우‡ 이일병
연세대학교 컴퓨터과학과, †서강대학교 전자공학과, ‡숙명여자대학교 전산학과

Document Understanding using Partial Matching Method

Yungcheol Byun○ Sungsoo Yoon Gyungwhan Kim† Yeong-Woo Choi‡ Yillbyung Lee
Dept. of Computer Science, Yonsei University
†Dept. of Electronic Engineering, Sogang University
‡Dept. of Computer Science, Sookmyung Women's University

요 약

여러 가지 유형의 서식 문서를 자동으로 처리하려면 서식을 이해하는데 필요한 항목 영상을 추출하기에 앞서 서식을 분류(classification)해야 한다. 서식을 분류함에 있어서 서식 영상 전체를 다룰 경우 상당한 시간이 걸릴 수 있다. 왜냐하면 일반적으로 서식 문서 영상의 크기는 일반 문자 영상에 비해 상당히 클 뿐만 아니라 대상 서식 문서의 유형도 많아질 수 있기 때문이다. 본 연구에서는 이러한 문제를 해결하기 위한 방법으로서 DP 매칭에 의한 부분 매칭 방법을 제안하고자 한다. 실험 결과, 제안하는 방법은 서식 문서의 전체가 아닌 일부 영역만을 비교함으로써 인식 시간과 인식률 면에서 서식 문서를 효과적으로 처리할 수 있었다.

1 서 론

1980년 이후 활발히 수행되어 온 서식 문서 처리에 관한 연구는 크게 두 갈래로 나뉘어 진다. 서식 문서 상에 존재하는 데이터를 자동으로 데이터베이스에 입력하고자 하는 연구와 전자 문서 시스템을 위하여 서식 문서를 컴퓨터 상의 문서로 자동으로 변환하고자 하는 연구가 그것이다. 전자는 사무 자동화의 관점에서 수행되었고 후자는 전자 문서의 자동 생성이라는 관점에서 수행되었다. 본 연구는 전자에 관한 연구로서, 서식 문서 상에 있는 데이터를 인식하여 데이터베이스로 자동으로 입력하기 위한 연구이다. 데이터를 자동으로 입력하기 위해서는 서식 문서 상에 있는 항목 영상 중 서식 문서를 해석하는데 필요한 필드 항목 영상을 추출한 후 OCR을 수행해야 한다. 처리해야 할 서식 문서의 유형이 한 가지일 경우 비교적 간단한 알고리즘을 이용하더라도 서식 문서를 처리할 수 있지만 여러 가지일 경우 보다 복잡한 알고리즘이 필요할 뿐만 아니라 항목 추출에 앞서 입력 서식을 분류하는 과정이 선행되어야 한다.

다양한 유형의 특징을 추출하여 이용함으로써 여러 가지 유형의 서식 문서를 분류하고자 하는 연구가 수행되었으며 긍정적인 결과를 얻었다[1]-[11]. 하지만 대부분의 연구는 서식 문서 전체 영역에서 서식 분류를 위한 특징을 추출함으로써 서식 문서 영상의 크기가 클 경우 특징 추출 시간이 많은 걸린다는 문제점을 가지고 있다. 특히 처리해야 할 서식 문서의 유형이 증가할수록 처리 시간은 증가하게 되며, 따라서 응용가능한(applicable) 서식 처리 시스템을 개발하기 위해서는 반드시 이를 해결해야 한다. 이러한 문제점은 서식 문서를 분류함에 있어서 서식 문서의 모든 부분을 동등하게 다루기 때문에 발생한다.

본 연구에서는 이러한 문제점을 해결하기 위한 방법으로서 부분 매칭에 의한 서식 분류 방법을 제안하고자 한다. 이를 위해 다음 장에서는 특징 추출 및 서식 분류에 대해 설명하고 3장에서는 부분 매칭에 의한 서식 분류 방법에 대해 설명한다. 그리고 4장에서 실험 결과에 대해 설명한 후 결론 및 개선점에 대해서는 5장에서 설명하고자 한다.

*본 연구는 정보통신연구진흥원 과제의 일부로 이루어졌음

2 특징 추출 및 서식 분류

2.1 처리하고자 하는 문서

본 연구에서 처리하고자 하는 문서는 지형적 구조(geometric layout structure)를 가지고 있는 서식 문서이다. 특히, 은행 신용카드 매출 전표, 은행 입출금표, 지로 용지 등과 같이 선분이 중요한 역할을 수행하는 문서로서, 예를 들면 (그림 1)과 같다. 이러한 서식 문서의 구조는 서식 문서를 구성하는 필드 항목 영상, 혹은 픽셀의 분포에 의해 형성되며 선분은 서식 문서의 구조를 명확히 해주는 역할을 한다.

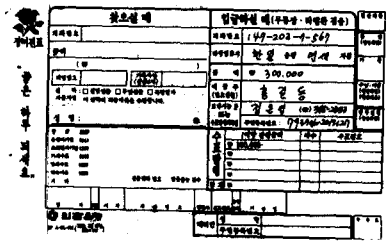

그림 1: 은행 입출금 전표 서식 문서의 예

일반적으로, 위와 같은 서식 문서에 항목 이외에 선분을 입력하는 경우는 거의 없으므로 서식 문서를 분류함에 있어서 선분 정보를 이용하는 것도 바람직하다고 볼 수 있다. 한편, 서식 문서 영상 전체에 대해 특징을 추출하여 분류할 경우 발생하는 문제점을 극복하기 위하여, 서식 문서 영상 전체가 아닌 일부 영역에 대해서만 특징을 추출하여 분류하도록 한다. 서식 분류시 서식 문서의 구조 정보를 이용하여 서식 문서를 분류할 것이므로 처리하고자 하는 서식 문서들 간에 서식 구조가 많이 차이나는 곳을 중심으로 비교하는 것이 바람직하다. 특히 서식의 구조를 명확히 하는 그래픽 요소가 선분이므로 서식 문서의 모든 부분

을 동등하게 다루기 보다는 서식 문서 간 선분에 의한 서식 구조가 많이 차이나는 곳을 위주로 비교함으로써 보다 효과적으로 서식을 분류할 수 있다.

2.2 서식 분류 관련 연구

서식 분류에 있어서 주요 이슈는 어떤 특징을 이용할 것인가와 그러한 특징을 입력으로 하여 어떤 인식기를 이용할 것인가에 관한 것이다. 특징 추출 과정은 입력 데이터의 양(차원)을 줄임으로써 인식기의 부담을 줄여주고 결과적으로 인식기의 성능을 높여준다. 즉, 인식률을 높이고 인식 시간을 줄일 수 있다.

서식을 분류하기 위해 사용되는 특징은 크게 물리적(physical) 특징과 논리적(logical) 특징으로 나눌 수 있다. 전자의 대표적인 예가 바로 픽셀의 물리적인 정보를 이용하는 픽셀 매칭인데, 이것 이외에도 문서의 크기(size)와 서식 상의 픽셀 밀도(density) 정보, 그리고 선분의 절대적 위치 정보 등이 있을 수 있다. 후자의 예로는 선분 간의 상대적인 위치 정보와 그러한 정보를 기반으로 한 서식 문서의 구조, 그리고 항목 간의 관계 정보와 서식 문서를 구성하는 구성 요소 간의 관계 정보가 있을 수 있다.

서식 문서 상의 수평/수직 선분은 텍스트와 비교적 쉽게 구별된다. 따라서 선분의 위치, 두께, 길이 등에 대한 정보를 추출한 후 DTW와 퍼지 매칭 방법 등을 이용하여 서식을 분류하고자 하는 연구가 수행되었다[1][6]. [2][3]에서는 입력 서식 영상을 $n \times n$ 영역으로 나눈 후 각 영역 내에서 선분 교차 특징을 추출한 후 특징 벡터를 구성하여 서식을 분류하였다. 한편, [4]에서는 테이블 내의 문자열 간의 인접 관계를 DP 매칭 방법으로 알아내어 테이블의 구조를 인식하는 방법을 제안하였고, [7]에서는 일반 프로젝션의 장점을 살리고 단점을 보완하는 스트립(strip) 프로젝션 방법을 이용하여 서식의 구조를 인식하는 방법을 제안하였다. 이 두 연구에서는 비록 서식 분류에 관한 언급은 없었지만 서식 구조 정보를 이용함으로써 서식을 분류할 수 있다.

[8]에서는 서식 문서 상의 모든 필드 항목이 수직/수평 선분으로 형성되도록 테이블을 구성한 후 필드의 좌상단 점을 이용하여 이진 트리를 구성하여 항목들 간의 관계를 표현함으로써 항목을 인식하는 방법을 제안하였다. 항목 간의 관계를 표현한 트리는 서식을 분류하는데 이용될 수 있다. [9]는 필드 항목에 해당되는 셀을 찾아 셀의 중심점을 비교함으로써 서식을 분류하는 방법을 제안하였다. 한편, 서식의 구조가 다르더라도 동일한 문서가 존재할 수 있다. 가령 서식의 물리적인 구조는 다르지만 항목 간의 인접 관계가 동일한 경우로서, 이러한 서식 문서 상의 항목을 인식하기 위하여 항목 간의 관계를 그래프로 표현하는 방법에 관한 연구가 있었다[10]. [11]에서는 픽셀 밀도를 입력으로 하는 MLP와 k-NN 인식기, 서식의 내용(contents)을 계층적으로 표현한 구조적 특징을 입력받아 트리 비교 방법으로 분류하는 구조적 분류기 등 세가지 분류기에 대해 설명하였다.

2.3 제안하는 방법

앞서 살펴 본 방법들은 서식 문서를 분류하기 위하여 서식 문서 전체를 동등하게 다루어 특징을 추출하였다. 이런 방법은 서식 문서를 유일하게 분류할 수 있다는 장점은 있으나 서식 분류시 특징 추출 시간이 많이 걸린다는 단점이 있다. 이는 응용가능한(applicable) 서식 문서 처리 시스템을 개발하기 위해서는 반드시 해결해야 할 문제로서, 이를 해결하기 위해 본 연구에서는 DP 매칭을 이용하여 매칭 영역을 결정한 후 해당 영역에 대해서만 특징을 추출하여 매칭을 수행하는 부분 매칭 방법을 제안한다.

3 부분 매칭에 의한 서식 분류

3.1 DP 매칭을 이용한 영역 결정

채워진 서식 문서의 경우 채워진 항목과 기존의(pre-printed) 텍스트를 구별하기란 그리 쉽지만은 않다. 하지만 잡음이 아닌 선분을 직접 입력하는 경우는 거의 없으므로 서식 분류를 위한 특징으로서 선분 정보를 이용하는 것은 바람직하다. 물론 문서 상의 항목이 주로 선분에 의해 결정되는 테이블 형식의 서식 문서를 처리할 경우 그렇다. 따라서 매칭할 영역을 결정함에 있어서 선분 정보를 이용하는 것이 바람직하다고 볼 수 있다.

입력 서식 문서를 등록된 서식 중의 하나로 효과적으로 분류하려면 등록된 서식들 간에 구조적 차이가 많이 있는 부분을 매칭 영역으로 결정해야 하는데, 이 경우 다음 사항을 고려해야 한다. 즉, 선분과 유사한 잡음이 추가되는 경우와 기존의 선분이 사라지는 경우, 선분이 끊기는 경우, 그리고 선분의 일부분이 사라지는 경우가 그것이다. 이러한 문제를 적절히 처리할 수 있도록 하기 위해 본 연구에서는 다음과 같은 DP 매칭에 의한 영역 결정 방법을 제안한다. 우선, (1)등록하고자 하는 서식 문서 영상을 $n \times m$ 영역으로 나누고 (2)각 영역에서 수평 선분과 수직 선분을 추출한다. 그리고 (3)수평 선분과 수직 선분 각각에 대해 선분 간의 거리를 구하여 특징 벡터를 구성한다. (4)등록하고자 하는 K개의 서식 문서에 대해 위의 연산을 수행한다. 이렇게 함으로써 등록하고자 하는 K 개의 서식 문서 각각에 대해 $n \times m$ 개의 특징 벡터를 추출할 수 있다. (5)이제 K 개의 서식 문서로부터 추출한 $n \times m$ 개의 특징 벡터에서 서로 대응되는 위치에 있는 2개의 특징 벡터 간의 유사도를 구한다. 이를 위해 다음과 같은 DP 매칭 방법을 이용하여 가중치 그래프(weighted graph)를 구한다.

$$g(i,j) = \min \left\{ \begin{array}{c} g(i+1,j) + C \\ g(i+1,j+1) + d(i,j) \\ g(i,j+1) + C \end{array} \right\}$$

i와 j는 가중치 그래프를 이차원 배열로 표현할 경우 행과 열 인덱스에 해당된다. 따라서 두 특징 벡터에 있는 요소의 수가 각각 M과 N개일 경우 $1 \le i \le M$, $1 \le j \le N$을 만족한다. 만일, 등록할 서식이 2개이고 각각의 서식 문서로부터 추출한 $n \times m$개의 특징 벡터 중 서로 대응되는 위치에 $(a_1, a_2, \ldots, a_M)$과 $(b_1, b_2, \ldots, b_N)$이라는 특징 벡터가 있다고 가정할 경우 $d(i,j)$는 다음과 같이 정의된다.

$$d(i,j) = |a_i - b_j|$$

가중치 그래프를 구한 다음, (6)가중치 그래프에서 다음을 만족하는 경로 $k(1), k(2), \ldots, k(Q)$를 찾는다. 이는 $(0,0)$ 위치의 노드와 (M,N) 위치의 노드를 잇는 최소 경로로서, 최소의 패널티(penalty)를 갖는 경로이다.

$$\min \left(\sum_{n=1}^{Q} w(n) \right)$$

이 경우 $w(n)$은 k_n 노드의 가중치 값을 반환하는 함수이다. 이처럼 $n \times m$개의 특징 벡터 쌍에 대해 DP 매칭을 수행한 후 패널티를 구한다. 그런 다음 (7)패널티가 큰 영역을 위주로 매칭 영역을 결정한다. 이는 결국 선분에 의한 구조적 특징이 많이 차이나는 영역을 비교함으로써 서식을 분류하겠다는 의미이다. 매칭 영역의 크기가 클 경우 특징 추출 시간은 증가하나 서식 문서 인식률은 높아지며, 반대로, 매칭 영역의 크기가 작을 경우에는 특징 추출 시간은 감소하나 문서 인식률은 낮아질 수 있다.

3.2 매칭 영역에 관한 지식

앞서 설명한 DP 매칭 방법을 이용하여 매칭할 영역을 결정한 후 매칭 영역에 관한 정보를 서식 모형으로 등록한다. 모형 기반 방

법에 대해서는 [12]를 참고하기 바란다. 매칭 영역에 관한 정보는 다음과 같은 FA 지식으로 기술하여 저장한다.

$$\mathsf{FA} = (L, (l, t, r, b))^{+}$$

이 경우 L은 라인을 의미한다. (l, t, r, b)는 서식 상의 매칭 영역을 의미하고 '+'는 한 번 이상 기술될 수 있음을 의미한다. 매칭 영역의 크기는 (l, t, r, b)로 표현되며, 매칭 영역의 수는 FA 지식을 구성하는 스크립트의 수로 표현된다. 물론 이러한 매칭 영역에 관한 지식은 서식 등록 과정에서 추출되어 모형의 일부분으로 등록되며, 서식 처리 과정에서 서식 분류를 위해 사용된다.

3.3 서식 분류

입력 서식 문서를 등록되어 있는 서식 문서 중의 하나로 분류하려면, 우선 입력 서식 문서에서 FA 지식으로 기술된 매칭 영역에 대해 수평/수직 선분을 추출한 후 선분 간의 거리 정보를 요소로 갖는 특징 벡터를 구성한다. 그런 다음 모형으로 등록되어 있는 K 개의 모형, 더 장확히 말하자면 각 모형의 FF 지식[12]과 DP 매칭을 수행함으로써 서식을 분류한다. DP 매칭 방법에 의해 서식 분류 문제는 가중치 그래프에서 패털티가 가장 적은 최적의 경로를 탐색하는 문제로 간주할 수 있다. 즉, 입력 서식은 패널티가 가장 작은 FF 지식을 모형으로 갖는 서식으로 분류할 수 있다. 서식을 분류한 후 모형으로 등록되어 있는 ITA 지식[12]을 이용하여 항목을 추출한다.

4 실험 결과

펜티엄 233 MMX 상에서 C++를 이용하여 앞서 제안한 방법을 구현하였으며, 7가지 유형의 시중 은행 입출금표 서식을 이용하여 제안한 방법의 성능을 테스트하였다. 우선, 각 유형의 채워지지 않은 서식 문서를 이용하여 7개의 영상을 획득한 후 서식 등록에 사용하였다. 그리고 채워진 서식을 이용하여 70개의 영상을 획득한 후 서식 분류 및 항목 추출 실험을 수행하였다. 두 경우 모두 200 dpi로 문서를 스캔하였으며, 이 경우 입출금표 서식 영상의 평균 크기는 1520×1070 픽셀이었다.

매칭 영역의 위치와 크기, 그리고 개수는 등록되어 있는 서식을 효과적으로, 그리고 유일하게 분류할 수 있도록 정의하는 것이 바람직하나 서식 문서의 크기와 유형이 다양하기 때문에 이러한 파라미터를 결정하기는 쉽지만은 않아 보인다. 따라서 현재로서는 매칭 영역 결정시 입력 서식을 9×6 영역으로 나누어 특징을 추출하였으며, 매칭 영역은 $54(9 \times 6)$개의 영역 중에 DP 매칭 결과 패널티가 가장 큰 상위 5개의 영역을 선택하였다. 파라미터 결정 방법은 향후 과제로 남기고자 한다.

항목 추출을 위해 각 서식 문서에서 이름, 금액, 비밀 번호 등 5-6 개의 항목을 등록하였다. 7개의 서식 모형을 등록한 후 입력 서식으로 부터 추출한 특징 벡터와 모형으로 등록되어 있는 특징 벡터에 대해 DP 매칭을 수행하여 서식을 분류하는데 걸린 시간은 평균 0.52초였다. 이 경우 잘못 분류하거나 분류하지 못한 경우는 없었다. 서식을 분류한 후 등록된 항목을 추출하는데 걸린 시간은 평균 0.32초였다.

5 결 론

서식 문서 처리시 서식 문서의 유형이 증가할수록 계산 시간이 늘어난다는 것은 당연하다. 그리고 서식 문서의 영상은 개별 문자 영상의 크기에 비해 상당히 크기 때문에 서식 문서 분류시 문서 전체를 액세스할 경우 서식 처리 시간은 심각한 문제로 대두될 수 있다. 예를 들어, 200 dpi로 전표와 은행 입출금표 등의 서식 문서를 스캔할 경우 문서 영상의 크기는 1000×1000 픽셀을 넘는 경우가 대부분이다. 이는 응용가능한 문서 처리 시스템을 개발하기 위해서는 반드시 고려해야 할 문제이다.

본 연구에서는 여러 가지 유형의 서식 문서 영상을 효과적으로 분류하고 서식 문서 해석에 필요한 항목 영상을 추출하기 위하여 DP 매칭을 이용한 부분 매칭 방법을 제안하였다. 이 방법은 서식 문서 분류시 문서 영상 전체를 비교하는 것이 아니라 구조적 특징의 차가 많이 발생하는 영역을 중심으로 비교함으로써 계산 시간을 줄일 수 있을 뿐만 아니라 양질의 특징을 추출함으로써 서식 인식률도 높일 수 있다. 하지만 앞서 실험한 방법에 대해 개선해야 할 점도 있다. 우선 파라미터를 결정하기 위한 보다 설득력있는 방법이 필요하다. 예를 들어, 매칭 영역의 수와 매칭 영역의 크기가 그것이다. 또한 선분에 의한 서식 구조 뿐만 아니라 서식 분류시 이용할 수 있는 유용한 특징에 관한 연구도 필요하다고 본다. 더 나아가, 비록 서식 구조가 같더라도 내용(contents)이 다름으로 인해 서로 다른 유형의 문서가 존재할 수 있으므로 그러한 서식 문서를 처리하기 위한 방법이 필요하다. 현재 그러한 것과 관련된 연구가 진행중에 있다.

참고 문헌

[1] Casey, R. G., D. R. Ferguson, K. Mohiuddin and E. Walach, "Intelligent forms processing system," *MVA*, Vol. 5, pp.511-529, 1992.

[2] Taylor, S. L., R. Fritzson and J. A. Pastor, "Extraction of data from preprinted forms," *MVA*, Vol. 5, pp.211-222, 1992.

[3] Lam, S. W., L. Javanbakht, and S. N. Srihari, "Anatomy of a form reader," *Proc. ICDAR*, pp.506-509, 1993.

[4] Yuki Hirayama, "A Method for Table Structure Analysis using DP Matching," *ICDAR*, pp.583-586, 1995.

[5] Watanabe, T. 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)

[6] J. Mao, M. Abayan, K. Mohiuddin, "A Model-Based Form Processing Sub-System," *ICDAR*, pp.691-695, 1996.

[7] Jiun-Lin Chen, Hsi-Jian Lee, "A Novel Form Structure Extraction Method Using Strip Projection," *ICDAR*, pp.823-827, 1996.

[8] Tsuneo SOBUE, Toyohide WATANABE, "Identification of Item Fields in Table-form Documents with/without Line Segments," *MVA*, pp.522-525, 1996.

[9] Shigeyoshi Shimotsuji, Mieko Asano, "Form Identification based on Cell Structure," *ICDAR*, pp.793-797, 1996.

[10] Y. Hirayama, "Analyzing Form Images by Using Line-Shared-Adjacent Cell Relations," *ICDAR*, pp.768-772, 1996.

[11] Pierre Heroux, Sebastien Diana, "Classification Method Study for Automatic Form Class Identification," *IWFHR*, pp.926-928, 1998.

[12] Yungcheol Byun, Yillbyung Lee, "Efficient Form Processing Methods for Various Kinds of Form Documents," *DAS*, pp.153-156, 1998.