세금계산서 상에서의 관심 데이터 추출

정 재 영, 유 돈 극
동양대학교 컴퓨터공학부

Field Data Extraction on Tax Form Image

Jung, Jae Young, Yoo, Don Keog
School of Computer Engineering, Dongyang Univ.

Abstract

본 논문에서는 세금 계산서 상에서의 관심 영역 및 관심 영역 내의 데이터를 추출하는 알고리즘을 제안한다. 먼저, 입력되는 세금 계산서 영상의 색상 정보를 이용하여 서식을 자동으로 추출한다. 추출된 서식 영상을 가지고 문서의 기울기 및 관심 대상 영역의 위치를 파악한 후, 원 영상에 대하여 관심영역을 추출한다. 관심영역에 대한 히스토그램을 분석하여 바탕 영역으로부터 인식 대상 데이터를 추출한다.

제안한 알고리즘을 다양한 화질의 세금 계산서 영상에 대하여 적용한 결과, 정확하게 관심 영역을 분할해내고 인식 대상 데이터를 추출할 수 있음을 보인다.

Ⅰ. 서론

현재까지 광학 문자 인식에 대한 연구는 상당한 수준까지 진척되어 이미 여러 시스템이 부분적으로 실용화되어있으며, 최근에는 필기 문자를 인식하고자 하는데까지 그 영역을 넓히고 있는 반면, 문서 영상의 구조 분석에 대한 연구는 상대적으로 미흡한 실정이다. 실제로 초기의 연구에서는 주로 여로가지 영상처리 기법들을 도입하여 주어진 문서 영상을 단순히 제목, 사진, 텍스트 등으로 구분되는 영역으로 나누는 데 초점을 맞추었다. 국내에서도 주로 신문이나 잡지, 교과서 등의 제한된 문서 영상을 대상으로 연구가 이루

어졌으며, 문서의 구조 분석을 통하여 기사를 추출하거나, 문자와 비문자를 분리 추출하는 등의 연구가 대다수를 차지하였다. 그 이후로 다양한 종류의 문서 영상 구조 분석 시스템에 관한 연구 결과들이 발표되었으나, 일반적인 문서 영상의 경우에는 처리해야 할 데이터 양이 너무 많고 종류도 다양하여 문서의 구조를 자동적으로 분석할 수 있는 시스템을 구축하는 것이 매우 어렵다는 인식 하에, 현재에는 문서의 형태가 일정한 서식을 갖추고 있는 문서 즉, 우편물, 수표, 입출금전표, 지로용지, 세금계산서 등과 같은 서식 문서 영상의 구조 분석에 관한 연구가 활발히 진행되고 있다.[1].

기존의 서식 문서 영상 구조 분석 방법들은 크게 모델 기반 방법[2~6]과 선 성분 기반 방법[7~10]으로 나뉘는데, 서식 문서에서 필요한 영역을 추출하기 위하여 문서 영상에서 추출하고자 하는 영역에 대한 정보를 미리 사용자가 직접 정의해주어야 하며, 처리하고자 하는 문서에 대한 많은 사전 지식과 경험적인 지식을 필요로 한다. 그러므로, 새로운 서식 문서를 처리하기 위해서는 시스템을 다시 설계하거나 필요한 영역을 다시 정의해주어야 하기 때문에 사용자의 입장에서는 매우 불편할 뿐만 아니라 부정확하므로 많은 비용과 시간의 낭비를 초래한다. 따라서, 새로운 문서에 대해 시스템이 보다 쉽게 적응할 수 있도록 추출하고자 하는 영역에 대한 정보를 사용자가 직접 정의하지 않고 그 정보를 시스템이 자동으로 획득하여 표현하며, 또한 문서에 대한 사전 지식 없이도 서식 문서에서 필요한 영역을 자동으로 추출하는 서식문서 구조 분석 시스템의 개발이 절실히 필요하다.

본 논문에서는 자동적으로 다양한 형태의 서식 문서를 형태에 따라 분류하고 서식 구조를 분석하여, 인식 대상 필드 영역으로부터 배경을 제거한 인식 대상 문자를 추출하는 서식 문서 영상 구조 분석 방법을 제안한다.

Ⅱ. 제안한 서식 문서 영상의 구조분석

일반적인 서식 문서 인식 시스템은 서식 문서를 입력받아 전처리 과정을 거쳐 서식 문서를 분석한 후, 관심 대상이 되는 문자를 분할한 후 인식하게 된다. 인식된 결과는 여러 가지 부가적인 정보를 이용하여 후처리 과정을 거쳐 검증된다.

본 논문에서는 서식 문서 인식을 위한 효율적인 전처리 방법과 서식 문서 분석을 통한 정보 문자 추출 방법을 제안한다. 전처리 과정에서는 효과적인 잡음 제거, 영상 형식 변환 기술, 기울기 검출 및 보정 기술[11] 등이 연구되고, 서식 문서 분석 과정에서는 서식의 분류, 각 서식의 구조 분석, 인식 대상 필드 영역 추출, 배경을 제거한 인식 대상 문자를 추출하는 방법을 제안한다. 제안한 서식 문서 분석 시스템의 전체적인 과정을 보이면 그림 1과 같다.

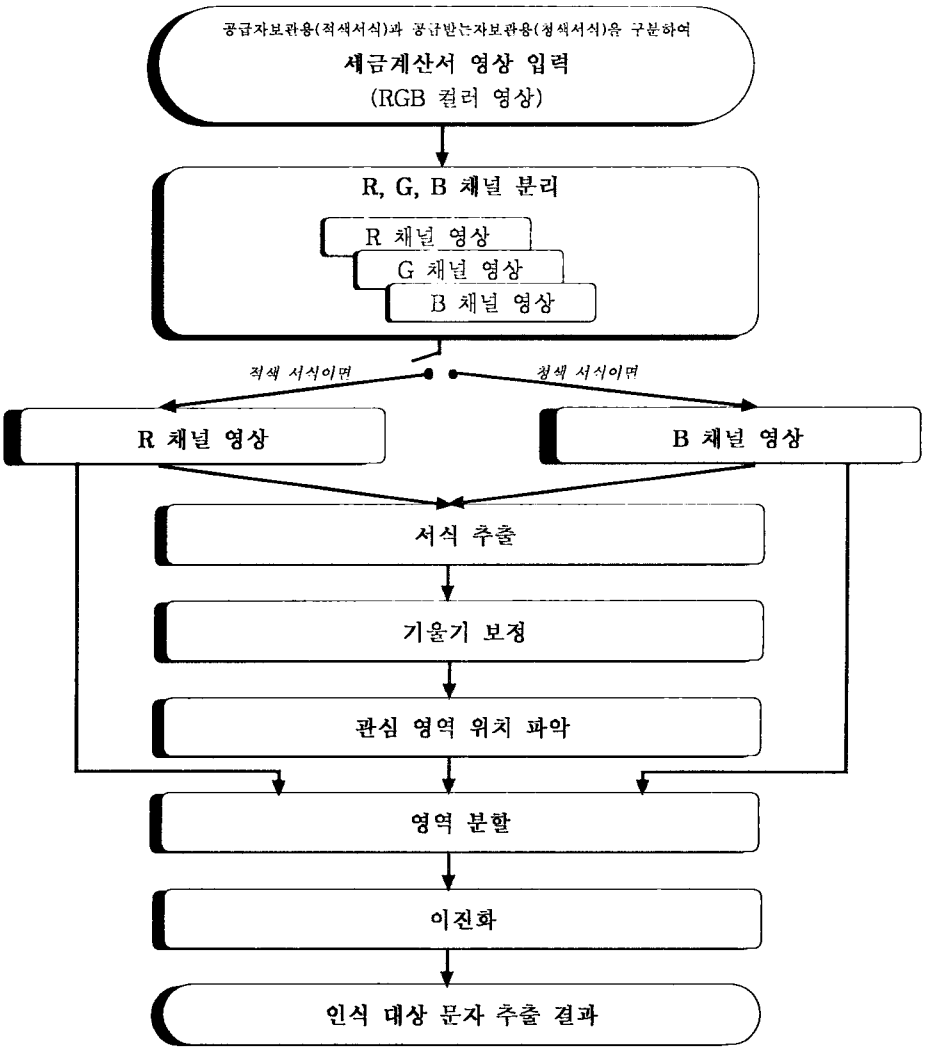


그림 1. 제안한 서식 문서 분석 시스템의 Overview

2.1 서식 추출

세금계산서를 공급자 보관용(적색배경)과 공급받는자 보관용(청색배경)으로 구분하여 RGB 컬러 영상으로 입력받아 그 서식을 추출하기 위하여 R, G, B 세 개의 채널로 분리한다. 그림 2는 공급받는자 보관용 세금계산서를 RGB 컬러 영상으로 입력받아 R, G, B 채널로 분리하였을 때의 B 채널 영상이다.

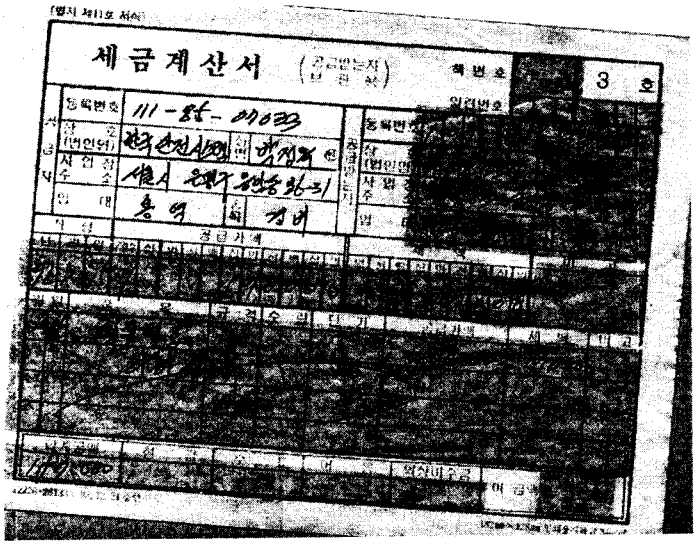


그림 2. 입력 영상에 대한 B 채널 영상

그림에서 보는 바와 같이 분리된 B 채널 영상은 크게 세 가지 서로 다른 영역으로 구분이 가능한데, 이는 서식, 바탕, 정보문자이다. 따라서, B 채널 영상으로부터 바탕과 정보문자를 제거한 순수한 서식만을 추출하기 위해서는 각 영역을 구분할 수 있는 특징을 파악하여야 한다.

본 논문에서는 각 영역을 구분하기 위한 특징으로 다음과 같은 사실을 경험적으로 파악하였다.

1) 바탕 영역의 각 화소는 R, G, B 각각의 값이 서로 유사하다. 이러한 특징은 그림 2와 같이 먹지로 인해 바탕 영역 내에 밝기가 크게 다른 부분들이 존재하는 경우나, 구겨진 문서가 입력되어 구겨진 부분에 흰 선이 생기는 경우나, 때가 묻어 있는 문서에도 적용이 가능하다.

2) 정보문자의 각 화소는 정보를 기입할 때 사용했던 펜의 색상이 그대로 나타난다. 만일 정보를 기입하는데 사용하는 펜의 색상을 서식 색상과 다른 것으로 사용하도록 제한할 경우(예를 들어 검정색 펜만을 사용), 서식 영역에 해당하는 화소들과 구분이 가능하다.

3) 서식 영역의 각 화소는 R, G, B 세 가지 색상 중에서 입력 용지에 따라 청색(공급받는자 보관용인 경우), 혹은 적색(공급자 보관용인 경우)이 다른 색상에 비해 강하게 나타난다. 따라서, 각 화소를 구성하는 R, G, B 값의 크기를 상대적으로 비교하여 그 화소가 서식 영역의 화소인지를 파악할 수 있다. 입력 영상의 청색(혹은 적색)의 절대적 크기를

고려하지 않고 이와 같이 상대적인 크기를 고려함으로써 용지 제작 회사에 따라 서로 다른 농도를 가지는 인쇄물에도 능동적인 적용이 가능하다. 일반적으로 공급받는자 보관용이든, 공급자 보관용이든 동일한 종류의 용지라 하더라도 용지 제작 회사에 따라 서로 다른 농도를 가지는 인쇄물이 가능하기 때문이다.

그림 2에 대하여 앞서 언급한 특징을 이용하여 서식을 추출하고, 각 화소를 동일한 밝기 값으로 표현한 결과를 보이면 그림 3과 같다. 그림 2의 B 채널 상에 존재하였던 바탕 영역이나 정보 문자 영역은 모두 제거되고 서식 영역만을 추출할 수 있게 된다.

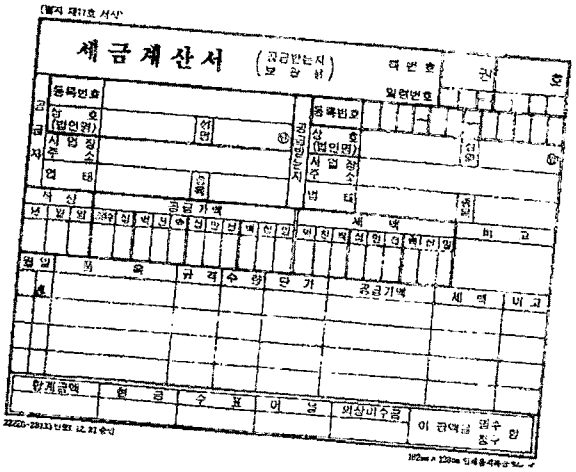


그림 3. 서식 추출 결과

2.2 문서 기울기 검출

본 절에서는 추출된 서식의 기울기를 계산하고, 일정 각도 이상 기울어진 경우 이를 보정하는 방법을 설명한다. 문서의 기울기를 보정하기 위한 연구는 크게 여백 탐색법, 투영에 의한 방법, 허프 변환을 이용하는 방법, 상관관계를 이용하는 방법 등으로 구분되는데, 본 논문에서는 정[11]의 공백행 추출에 의한 기울기 보정 방식을 이용하였으며, 기본적인 알고리즘은 표 1과 같다.

표 1. 기울기 추정 알고리즘

입력 : 기울어진 이진 문서 영상 출력 : 기울어짐이 보정된 이진 문서 영상 알고리즘 : 1. Dilation에 의한 화소 widening 2. 수직 샘플링에 의한 행간점 추출 2.1 전체영상에 대해 일정 간격으로 수직샘플링 수행 2.2 각 수직 단면으로부터 행간점 추출 3. 행간점 쌍간의 기울기 계산 4. 히스토그램 분석 => 최대치:기울어진 각도

이 방법의 경우, 기울기 검출 방식이 문서가 기울어진 정도에 무관하고, 도표나 적은 비율의 텍스트 영역을 차지하는 문서 등에 관계없이 효율적으로 문서 기울기 검출이 가능하다. 또한 허프 변환을 이용한 방법이나 퓨리에 변환을 이용한 방법과 같이 문서 내의 연결 화소를 찾는다거나 변환 함수를 적용하는 과정을 요구하지 않으므로 처리 시간이 빠르다.

그림 3의 서식 문서에 대하여 문서의 기울기가 보정된 결과를 보이면 그림 4와 같다.

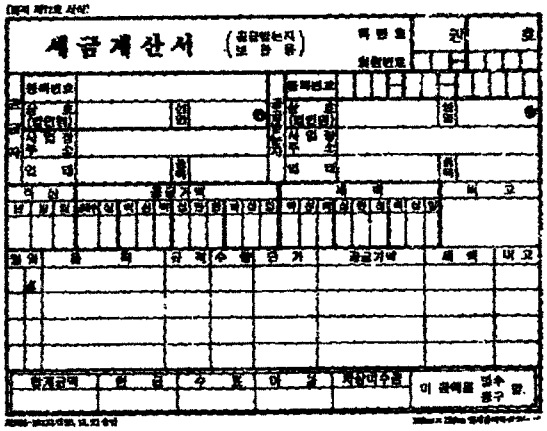


그림 4. 기울기가 보정된 서식 문서 영상

2.3 영역분할

본 절에서는 입력 영상으로부터 관심 영역(공급자의 등록번호, 공급받는자의 등록번호, 작성년도와 공급가액 및 세액)을 찾는 과정을 설명한다.

먼저, 영상 상단부에 있는 두 등록번호 영역을 검출하기 위하여, 투영법을 이용하여

영상 상단부(전체 높이의 1/2)에 존재하는 긴 수평선과 수직선을 검출하고, 이들의 위치 관계를 이용하여 두 등록번호 영역을 검출한다.

그림 4와 같이 기울기가 보정된 서식 문서로부터 긴 수평선을 추출하기 위하여 수평 방향으로 투영하여 검은 화소의 개수를 누적시킨다. 이 때, 일정 길이 이상으로 연속적으로 이어지는 검은 화소만을 고려한다. 이는 서식상에 존재하는 문자들은 가능한 한 배제하기 위함이다. 누적된 검은 화소의 개수가 전체 영상 폭의 일정 비율 이상인 수평선들을 추출한다. 추출된 수평선들이 연속적으로 인접한 경우, 이는 실제로 하나의 수평선이 두껍게 나타나는 경우이므로 세선화 과정을 거쳐야 한다. 본 논문에서는 이러한 세선화 과정을 손쉽게 처리하기 위하여 인접한 행들이 검출될 경우 그들의 중앙에 위치한 선을 계산하여 이용한다. 또한, 서식 문서 상단부에 존재하는 긴 수직선을 검출하기 위해서는 수직 방향으로 투영한 후 수평선 검출과 동일한 과정을 반복하게 된다. 이와 같이 추출된 수평 수직선의 위치를 이용하여 공급자의 등록번호와 공급받는자의 등록번호 영역 위치를 결정할 수 있게 된다.

작성 연월일, 공급가액, 세액 영역의 위치를 결정하기 위해서는 전체 서식 영상을 높이에 따라 3등분 하고, 이들 중 중앙부의 각 행을 조사하여 흑백의 교차 회수가 가장 많은 행을 검출한 후, 그 행과 가장 가까운 하위 두 개의 긴 수평선을 찾는다.

그림 5는 앞서 설명한 방식으로 서식 문서 영상에서 검출된 긴 행과 열의 위치를 검은 점으로 표시하여 보인 것이다.

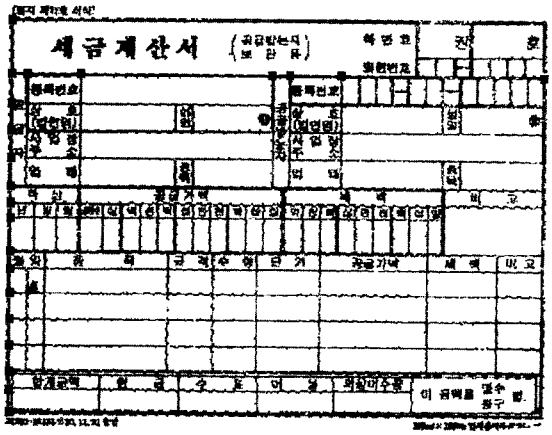


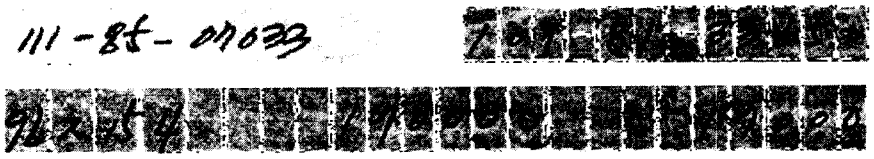
그림 5. 서식 문서 영상으로부터 추출된 수평선과 수직선의 위치

2.4 관심 데이터 추출

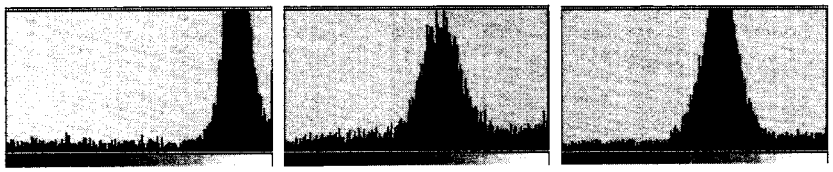
지금까지 서식 문서를 이용하여 문서의 기울기를 보정하고, 관심 영역의 위치를 검출

하는 과정을 설명하였다. 이제, 입력 영상으로부터 얻어진 B 채널 또는 R 채널 영상에서 관심 영역을 분할해 내고, 그 안에서 배경을 제외한 관심 데이터를 추출하는 과정이 필요하다.

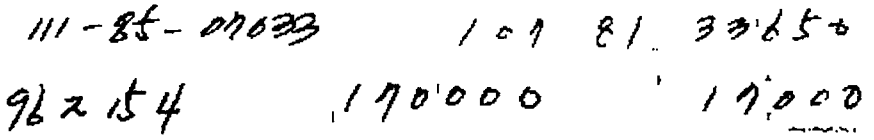
그림 6(a)는 그림 2와 같은 B 채널 영상에서 분할된 세 개의 관심영역으로 (b)는 각 영역에 대한 히스토그램 분포이다. 그림 (b)에서 알 수 있듯이 각 영역은 배경 부위에서 많은 화소를 가지게 되며, 그보다 밝기 값이 작은 화소들은 문자를 구성하는 화소들이다. 따라서 적절한 임계치를 이용하여 각 영역을 이진화할 경우 그림 (c)와 같이 배경을 제거한 문자 데이터만을 구할 수 있게 된다.



(a) 세 개의 관심영역



(b) 각 영역별 히스토그램



(c) 데이터 추출 결과

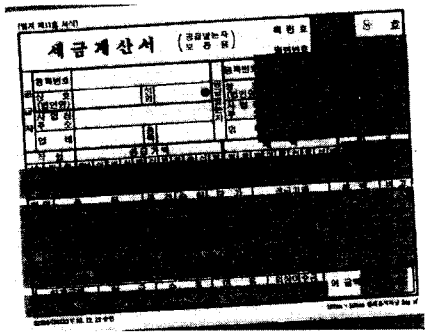
그림 6. 세 개의 관심영역과 그들의 히스토그램 및 추출된 문자

Ⅲ. 실험결과

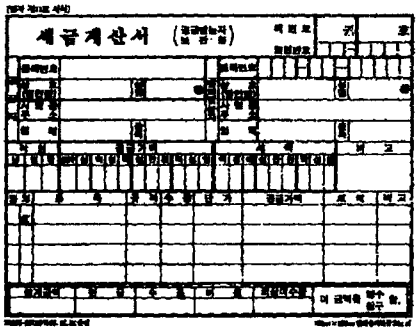
제안한 방식을 C 언어로 구현하여 100~300dpi로 읽혀진 다양한 유형의 RGB 컬러 세금계산서 영상에 적용하였다. 입력 영상은 적색 서식의 공급자보관용인지 청색 서식의 공급받는자 보관용인지가 구분되어 입력된다.

그림 7(a)는 먹지로 인하여 바탕이 글자색과 매우 유사한 기울어진 영상으로, 이에 대한 서식 추출 및 기울기 보정 영상은 그림 7(b)와 같다. 그림 (b)에서 알 수 있듯이 정확

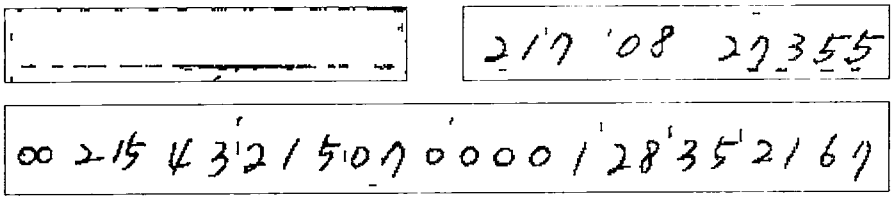
한 서식의 추출과 기울기의 보정이 이루어지고 있다. 그림 (c)는 원 영상으로부터 데이터 영역을 추출하고 바탕색을 제거한 결과로 정보의 손실없이 원하는 결과가 얻어짐을 알 수 있다.



(a) 원 영상



(b) 기울기가 보정된 서식 추출 영상

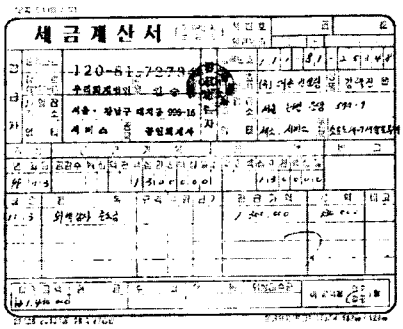


(c) 데이터 추출 결과

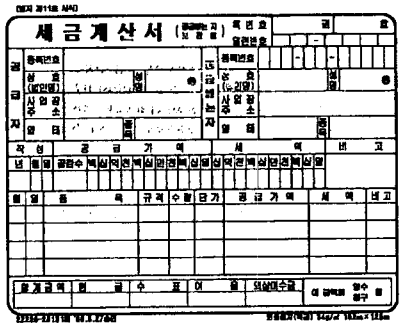
그림 7. 실험 영상 1

그림 8(a)는 데이터가 해당 영역에 걸쳐 있는 경우(공급자의 등록번호 부분)의 영상으로, 이에 대한 서식 추출 결과 영상은 그림 7(b)와 같다. 원 영상에 대한 기울기 계산 결과가 0도로 기울기 보정 과정은 수행되지 않는다. 본 실험에서는 처리 속도를 빠르게 하기 위하여 계산된 기울기의 절대치가 0.1도 미만일 때에는 기울기 보정 과정을 거치지 않도록 하였다. 그림 (c)는 원 영상으로부터 데이터 영역을 추출하고 바탕색을 제거한 결과로 정보의 손실없이 원하는 결과가 얻어짐을 알 수 있다.

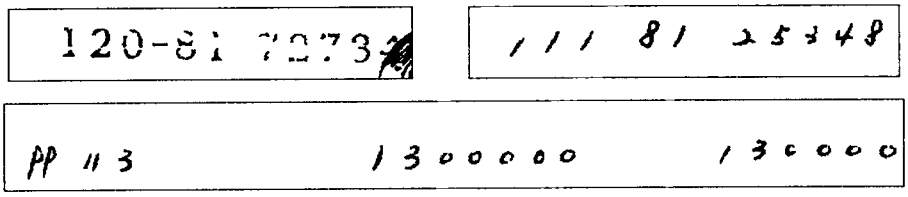
그림 9(a)는 입력된 숫자 데이터의 색상이 서식의 생상과 유사한 경우(공급자의 등록번호 부분)의 영상으로, 이에 대한 서식 추출 결과 영상은 그림 7(b)와 같다. 그림 (b)에서 알 수 있듯이 공급자의 등록번호가 서식의 색상과 유사하여 서식으로 오인되어 잘못 추출되고 있다. 그림 (c)는 원 영상으로부터 데이터 영역을 추출하고 바탕색을 제거한 결과로 공급자의 등록번호를 제대로 추출하지 못하고 있다.



(a) 원 영상

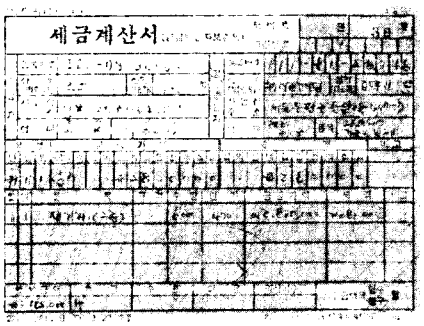


(b) 기울기가 보정된 서식 추출 영상

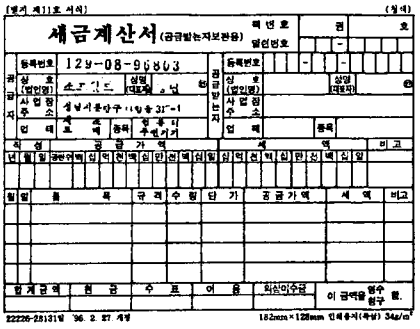


(c) 데이터 추출 결과

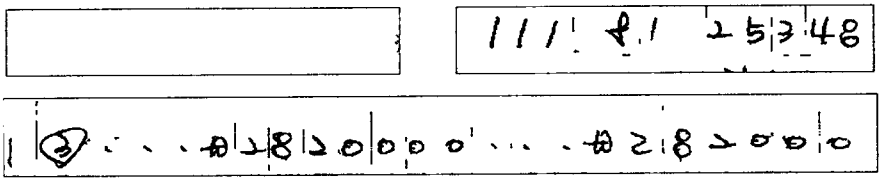
그림 8. 실험 영상 2



(a) 원 영상



(b) 기울기가 보정된 서식 추출 영상



(c) 데이터 추출 결과

그림 9. 실험 영상 3

Ⅳ. 결론

본 논문에서는 서식 문서 인식을 위한 전처리 과정으로, 입력되는 문서의 서식을 추출하고 인식 대상이 되는 숫자 데이터를 추출하는 알고리즘을 제안하였다. 제안한 방식에서 추출하고자 하는 데이터 영역의 위치 결정은 자동 추출된 서식으로부터 이루어지며, 아무런 사전 정보를 요구하지 않는다는 장점이 있다. 추후 연구과제로는 서식과 동일한 색상으로 숫자 데이터가 기입되었을 경우, 이를 서식으로 오인하지 않고 이를 정확히 추출하기 위한 연구가 필요하다.

참 고 문 헌

[1] 이성환, 문자인식 : 이론과 실제, 홍릉과학출판사, 1994년 4월.

[2] H. Fujisawa and Y. Nakano, "A top-down approach for the analysis of document images," Proc. Workshop on Syntactic and Structural Pattern Recognition, pp.113-122, Jun., 1990.

[3] J. Higashino, H. Fujisawa, Y. Nakano, and Ejin, "A knowledge-based segmentation method for document understanding," Proc. 8th Int. Conf. on Pattern Recognition, pp.745-748, Oct., 1986.

[4] C. D. Yan, Y. Y. Tang, and C. Y. Suen, "Form understanding system based on form description language," Proc. 1th Int. Conf. on Document Analysis and Recognition, pp.283-293, Sep., 1991.

[5] S. W. Lam, L. Javanbakht, and S. N. Srihari, "Anatomy of a form reader," Proc. 2th Int. Conf. on Document Analysis and Recognition, pp.506-509, Oct., 1993.

[6] J. Yuan, L. Xu, and C. Y. Suen, "Form items extraction by model matching," Proc. 1th Int. Conf. on Document Analysis and Recognition, pp.210-218, Sep., 1991.

[7] S. Chandran and R. Fasturi, "Structural recognition of tabulated data," Proc. 2th Int. Conf. on Document Analysis and Recognition, pp.516-519, Oct., 1993.

[8] T. Watanabe, H. Naruse, Q. Luo, and N. Sugie, "Structure analysis of form documents on the basis of the recognition of vertical and horizontal line segments," Proc. 1th Int. Conf. on Document Analysis and Recognition, pp.638-646, Sep., 1991.

[9] T. Watanabe, Q. Luo, and N. Sugie, " Toward a practical document understanding of table-form documents: Its framework and knowledge representation," Proc. 2th Int. Conf. on Document Analysis and Recognition, pp.510-515, Oct., 1993.

[10] D. Wang and S. N. Srihari, "Analysis of form images," Proc. 1th Int. Conf. on Document Analysis and Recognition, pp.181-191, Sep., 1991.

[11] 정재영, 김문현, "행간점 추출에 의한 고속 문서 기울기 검출," 정보과학회, 제26권 제11호, pp. 1342-1349, 1999년 11월.