Journal of Digital Contents Society Vol. 15 No. 2 Apr. 2014(pp. 209-217)
http://dx.doi.org/10.9728/dcs.2014.15.2.209

표 서식 문서의 구조 분석을 위한 선분 에지 기반의 유형별 꼭짓점 검출

정재영*

요 약

표 서식을 활용하고 있는 수많은 문서들을 종류에 따라 자동으로 분류하거나, 서식에 기입된 정보를 서식과 분리하여 추출하는 기술은 매우 중요하게 활용된다. 이를 위해서는 표 서식 구조를 정확하게 파악하는 과정은 필수적이다. 본 논문에서는 표 서식 문서 영상에 대한 유형별 꼭짓점 검출 방법을 제안한다. 주요 처리 과정은 전처리, 에지 블록 검출, 선분 에지 블록 검출, 꼭짓점 검출 단계를 거친다. 추출된 꼭짓점들은 선분 에지들이 다양한 형태로 직교하는 교차점들로 9가지 유형으로 분류된다. 실험에서는 제안한 방법을 세금계산서, 거래명세표, 표를 포함하고 있는 일반 문서 등과 같은 몇 가지 형태의 영상에 적용하여 99% 이상의 유형별 꼭짓점 추출 성능 결과를 보인다. 서식 문서 내에서의 대부분의 꼭짓점들은 대칭 형태로 존재한다는 사실을 고려할 때, 꼭짓점의 유형, 선분 에지의 폭 및 그들의 위치 관계를 활용하여 서식의 구조 분석에 활용 가능하다.

키워드 : 꼭짓점 검출, 선분 에지, 기울어짐 보정, 표 서식 문서 영상

Line Edge-Based Type-Specific Corner Points Extraction for the Analysis of Table Form Document Structure

Jae-young Jung*

Abstract

It is very important to classify a lot of table-form documents into the same type of classes or to extract information filled in the template automatically. For these, it is necessary to accurately analyze table-form structure. This paper proposes an algorithm to extract corner points based on line edge segments and to classify the type of junction from table-form images. The algorithm preprocesses image through binarization, skew correction, deletion of isolated small area of black color because that they are probably generated by noises.. And then, it processes detections of edge block, line edges from a edge block, corner points. The extracted corner points are classified as 9 types of junction based on the combination of horizontal/vertical line edge segments in a block. The proposed method is applied to the several unconstraint document images such as tax form, transaction receipt, ordinary document containing tables, etc. The experimental results show that the performance of point detection is over 99%. Considering that almost corner points make a correspondence pair in the table, the information of type of corner and width of line may be useful to analyse the structure of table-form document.

Keywords : Corner Point Extraction, Line Edge, Skew Correction, Table-form Document Image

1. 서론

예금 전표, 신용카드 전표, 세금계산서, 거래명세표 등 서식 문서의 형태는 매우 다양하며, 광

※ 교신저자(Corresponding Author): Jae. Y. Jung
접수일:2014년 01월 15일, 수정일:2014년 02월 25일
완료일:2014년 04월 07일
* 동양대학교 컴퓨터정보전학과
Tel: +82-54-630-1055 , Fax: +82-54-630-1141
email: jyjung@dyu.ac.kr

▣ 본 연구는 동양대학교의 2013학년도 교내연구비 지원에 의해 수행되었음

범위한 분야에서 활용되고 있다[1~3]. 일반적으로 이러한 서식문서는 컴퓨터시스템 상에서 생성되고 인쇄되어 일상생활에서 추가 정보가 기입된 후, 전자 문서로 다시 변환되는 일련의 과정을 거치는데, 수작업으로 이들 서식을 분류하고 전자 문서로 변환하는 것은 매우 많은 시간과 노력을 요하는 작업이다. 따라서 이를 자동화하기 위한 서식 문서 구조 분석 시스템의 개발이 절실히 필요하다[4~5].

기존의 서식 문서 구조 분석에 대한 연구는 입력된 서식 문서 영상으로부터 수많은 서식 중에 특정 서식을 식별하거나, 사용자에 의해 서식에 추가로 기입된 정보를 추출하는 연구로 구분되는데, 많은 연구들이 서식 문서에서 추출하고자 하는 항목 영역에 대한 사전 정보를 사용자가 정의하여 주고 있다. 이는 사용자에게는 매우 불편한 과정일 뿐만 아니라, 새로운 형태의 서식 문서를 처리할 경우 시스템을 재설계해야 하는 단점이 있다.

대부분의 서식 문서 분석 시스템이 수행하는 주요 과정으로는 정보가 기입된 인쇄용지를 스캔하여 디지털 영상으로 변환하는 영상취득 과정, 잡음제거, 이진화, 세선화, 기울기 보정 등과 같이 이후 처리 단계에서의 성능을 높일 수 있도록 사전 처리 하는 전처리 과정, 특정 서식의 구조를 파악하고 셀의 위치와 구성을 특징벡터로 표현하는 구조 분석 과정, 서식 내부에 관심이 되는 셀의 위치를 찾고 해당 영역을 추출하는 관심영역 추출 과정, 관심 영역으로부터 기입된 정보를 추출하고 인식하는 데이터 해석 과정을 거친다.

본 논문에서는 서식 문서 분석 처리 과정에서 서식의 구조를 분석하기 위하여 선분 에지에 기반한 유형별 꼭짓점 검출 방법을 제안한다. 검출된 꼭짓점들은 표를 표현하는 직선 선분들이 교차하는 점들로 이들을 정확하게 검출해 내는 것은 영역 추출 및 데이터 분석과 같은 이후 처리 과정에서 보다 우수한 성능을 내기 위해 필수적이다. 제안한 알고리즘의 특징은 다음과 같다.

- 직선 선분으로 테두리를 가지는 표 서식에 대하여 적용 가능하다.
- 임의의 표 구조를 가지는 서식에 대하여 적용 가능하며 서식에 대한 사전 정보를 요구하지 않는다.
- 서식 영상의 해상도는 크게 영향을 받지 않아서 100dpi의 저해상도 영상에 대해서도 적용 가능하다.
- 서식 문서뿐만 아니라, 표를 포함하고 있는 일반 문서에도 적용 가능하여 문서 DB 상에서의 표 검색 작업에 활용 가능하다.
- 표상에서 꼭짓점이 잘 드러나는 필기로 작성된 서식 문서에도 적용 가능하다.

2장에서는 관련된 연구를 소개하며, 3장에서는 제안한 꼭짓점 검출 방법을 설명한다. 4장에서는 제안한 방식을 세금계산서, 표를 포함하고 있는 일반문서, 수작업으로 작성된 표문서 등에 적용한 결과를 보인다. 5장에서는 결론과 추후 과제를 남긴다.

2. 관련연구

기존의 서식 문서 영상 구조 분석 방법들은 크게 모델 기반 방법과 선 성분 기반 방법으로 나눌 수 있다.

모델 기반 방법에서는 사용자가 관심 영역 추출에 필요한 정보를 여러 가지 형태로 시스템에 제공하여 처리한다. 구체적으로 동일한 구조의 저해상도 샘플영상으로부터 망(mesh) 구조를 계산하여 서식 모델을 구하는 방법[6], 7가지 상자 유형을 정의하고 상자들의 집합으로 문서 구조를 분석하는 방법[7] 등이 있다. 그러나 이러한 방법들은 사용자의 개입을 요구함은 물론, 다양한 서식에 대하여 서로 다른 정보들을 제공해야 하는 단점이 있다.

선 성분 기반 방법은 대부분의 서식 문서가 사각형 형태의 직선 성분으로 구성되어 있다는 점에 착안하여 선분을 이용하거나 사각형의 모서리를 이용하여 처리하는 방법으로 구분된다. Watanabe[8]은 셀의 좌상단 꼭짓점을 이용하여 테이블을 표현하였다. 몇 개의 필터를 사용하여 수직선과 수평선을 추출하고, 이를 기반으로 좌상단 모서리를 검출한다. 그러나 이 방법에서는 좌상단 꼭짓점만을 교차점으로 활용하기 때문에 검출 오류에 대한 영향을 크게 받는다.

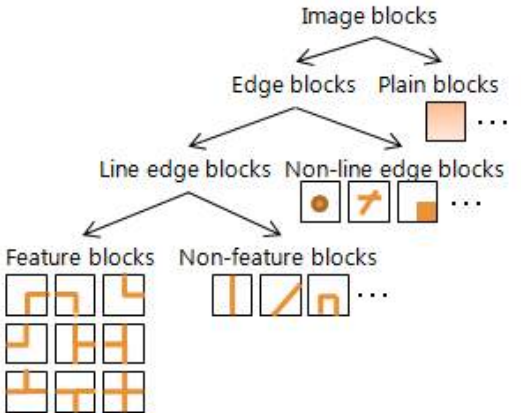
Taylor[9]는 선분의 폭이 1픽셀 두께라는 가정하에 9X9 크기의 필터를 이용하여 서식의 모서리에서 나타나는 4가지 유형의 교차점을 찾고, 그 다음 이를 조합하여 5가지 유형의 추가적인 교차점을 찾는다. 그러나 그의 방식에서는 문서의 종류에 따라 선분의 폭이 달라질 경우 필터를 변경하여 적용하여야 하는 문제를 안고 있다. Neves[10]는 Taylor[9]의 방법을 개선하기 위하여 필터 대신 모폴로지 침식(erosion) 연산을 수행하여 교차점을 찾고 있다. 그의 방법 역시 2단계 유형 검출 과정을 거침으로써 오류의 단계별 전파 영향을 크게 받고 있다. 특히, 잡음에 매우 민감하여 서식이 비어있으면서 단순한 형태의 고해상도 영상에 대해서는 좋은 결과를 보이고 있으나, 셀의 개수의 많아지고 정보가 기록된 문서 영상에 대해서는 검출 성능이 현저히 떨어진다.

3. 유형별 꼭짓점 검출

에지는 영상에서 차지하는 비율은 낮으면서도 영상을 이해하는데 중요한 정보를 내포하고 있다. 특히 표 서식 문서의 경우 서식을 표현하는 수평 수직 선분 에지 정보는 서식의 레이아웃을 판단하는데 매우 중요한 요소이다.

본 논문에서는 영상 위에 nXn 크기의 마스크를 한 화소씩 이동시켜 나가면서 블록을 추출한 후, 그 블록 내의 밝기 값 변화를 조사하여 에지 블록을 찾고, 그 블록의 경계선에 선분이 걸치는지를 조사하여 선분 에지 블록인지 아닌지를 판별한다. 선분 에지 블록에 대해서는 수평 수직 선분이 만나거나 교차하여 생기는 꼭짓점 존재 여부에 따라 특징 블록과 특징 블록이 아닌 것으로 구분한다. (그림 1)의 “Non-line edge” 블록의 예에서 보인 세 개의 에지 블록 중, 앞 쪽의 두 개는 블록의 경계선에 접하는 선분이 존재하지 않아 선분 에지 블록에서 배제되며, 마지막 블록은 블록의 각 경계면을 기준으로 선분이 블록 내부에 완전히 포함되지 않아서 비 선분 에지 블록으로 분류하였다. “Non-feature” 블록의 예에서 보인 블록 중 마지막 것은 꼭짓점이 존재하기는 하나, 블록의 크기가 비교적 작다는 것을 고려할 때 일반적인 표 서식에서 그와 같은 2개의 대칭형 꼭짓점들이 실제 존재한다기보다는 잡음, 서식 위에 기입된 정보 등에 의해 나타난 것으로 보고 비 특징 블록으로 분류하였다. 마지막으로, 추출된 특징 블록으로부터 꼭짓점을 검출하여 9가지 유형 중의 하나를 판별한다.

(그림 1) 영상 블록의 분류와 예



(Figure 1) Block classification and samples

3.1 전처리

효율적인 꼭짓점 검출을 위하여 입력되는 서식 문서 영상에 대하여 이진화, 기울기 보정, 잡음 제거와 같은 전처리 과정을 거친다.

다양한 밝기와 화질로 입력되는 그레이 영상을 이진 영상으로 변환하기 위해서는 어두운 영역과 밝은 영역을 구분하기 위한 임계값 설정이 중요한 문제인데, 입력된 영상의 상태를 판단하여 해당 영상에 적합한 임계값을 자동적으로 설정하는 것이 중요하다. 일반적인 문서 영상의 경우 이봉(bimodal) 형태의 히스토그램 분포를 가지는 데[11], 본 논문에서는 빠른 이진화를 위하여 히스토그램 상에서 나타나는 극값들의 변화를 활용하여 입력 영상에 따라 능동적으로 임계값을 구하였다. 이후, 이진 영상에 대하여 서식의 기울기를 계산하고, 일정 각도 이상 기울어진 경우 이를 보정한다. 문서의 기울기를 보정하기 위한 연구는 크게 여백 탐색법, 투영에 의한 방법, 허프 변환을 이용하는 방법, 상관관계를 이용하는 방법 등으로 구분되는데, 본 논문에서는 공백행 추출에 의한 기울기 보정 방식을 이용하였다[12]. 이 방법에서는 기울기 검출 방식이 문서가 기울어진 정도에 무관하고, 도표나 낮은 비

율의 텍스트 영역을 차지하는 문서 등에 관계없이 효율적으로 문서 기울기 검출이 가능하다. 또한 허프 변환을 이용한 방법이나 푸리에 변환을 이용한 방법과 같이 문서 내의 연결 화소를 찾는다거나 변환 함수를 적용하는 과정을 요구하지 않으므로 처리 시간이 빠르다. 그레이 영상을 이진 영상으로 변환하게 되면 작게 고립된 검은 점들이 나타나게 되는데, 이는 잡음 등의 영향으로 원 영상에서 존재하였던 작고 흐릿한 부분이 이진화의 결과로 분명하게 드러나기 때문이다. 본 논문에서는 4X4 크기의 마스크 내부에 완전히 포함되는 검은 점들은 잡음으로 판단하고 이를 제거하였다. 완전히 포함된다는 것은 마스크의 경계 영역은 모두 흰 점이고, 중앙의 2X2 영역에 검은 점이 존재하는 경우를 의미한다.

3.2 에지 블록 검출

에지 블록을 검출하기 위하여 블록 전체의 평균 밝기 값을 계산하고, 그 평균값을 기준으로 밝은 영역과 어두운 영역으로 구분하여 각 영역의 밝기 평균을 계산한다. 두 영역의 평균 밝기 값의 차이가 임계값 이상이면 에지블록으로 분류한다. 구체적인 알고리즘은 <표 1>과 같다.

<표 1> 에지 블록 검출 과정

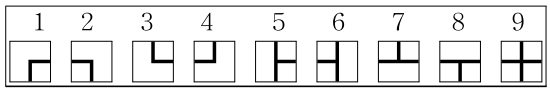
//$n \times n$: size of a block //Thr : threshold of bright gradient For all image pixel (i, j) 1. Formulate a $n \times n$ block, X, centered at (i, j) 2. Classify X into edge block or not. 2.1 Calculate the average intensity of all pixels in the block X, $\overline{X}$. 2.2 Calculate the average intensity of pixels greater than $\overline{X}$ in X, $\overline{X_h}$. 2.3 Calculate the average intensity of pixels smaller than $\overline{X}$ in X, $\overline{X_l}$. 2.4 if ($\overline{X_h} - \overline{X_l} > Thr$), X : edge block.

<Table 1> Edge block detection

3.3 선분 에지 검출

검출된 에지 블록들은 그 유형이 매우 다양하다. 이 중, 에지가 수평/수직 선분의 형태로 접하거나 교차하여 만들어질 수 있는 꼭짓점을 (그림 2)와 같이 9가지 유형으로 분류하였는데, 이들은 표에서 나타날 수 있는 모든 유형을 포함하며 서식 문서를 분석함에 있어 중요한 특징에 해당한다.

(그림 2) 9가지 유형의 특징 블록



(Figure 2) feature blocks of 9 types

본 절에서는 에지 블록에 대하여 선분 에지를 포함하는지, 선분 에지를 포함한다면 꼭짓점을 포함하는 특징 블록인지, 그 유형은 어떠한지를 판별하기 위한 과정을 설명한다.

에지 블록(nXn)에 있는 상하좌우 4개의 경계면(각 면은 n차원 벡터)을 각각 독립적으로 조사하여 이웃 화소끼리 밝기 변화가 백->흑 또는 흑->백으로 바뀌는 위치를 파악하여 wb 또는 bw로 설정한다. 각 경계면에서 wb, bw 쌍이 순서대로 나타날 때 블록 안쪽으로 하나의 선분이 진입하는 것으로 판단할 수 있고, 그 선분의 폭($Width$)은 |wb-bw|이며, 세선화(thinning)할 경우 선분의 시점은 [wb, bw] 중의 한 점이 된다. 해당 선분의 길이를 조사하기 위하여 폭 범위 [wb, bw]의 각 점에서 블록 내부로 연결되는 직선상에서 검은 화소가 연속적으로 나타나는 개수 $Black\,Run$ 을 구한다. 이 때, 잡음의 영향을 고려하여 연속된 검은 화소 중간에 길이가 1인 흰 화소는 연결된 것으로 간주하여 처리한다. 이후 식(1)과 같이 그 중 최댓값을 선분의 길이($Length$)로 결정하고, 그 선분의 길이가 임계값(실험에서는 마스크 크기의 50%로 설정)보다 크면 그 때의 블록 경계선 상의 위치를 선분의 시점(Pos)으로 정의하였다. 이는 블록에 포함된 선분의 길이가 너무 작을 경우, 선분에 대한 정보가 부족하여 기입된 문자 등에서 잘못 검출될 가능성을 배제하기 위함이다.

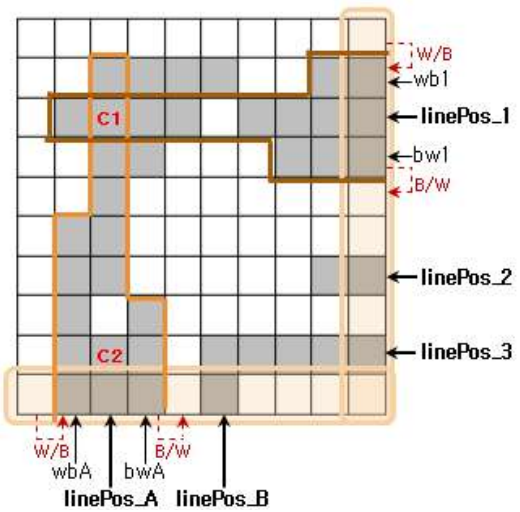
$$Length = \max_{p=wb}^{bw}\{BlackRun(p)\}$$
$$Pos = p,\ \text{if}\ Length > Threshold \quad (1)$$

최종적으로, 선분 에지는 식(2)와 같이 정의된다.

$$Line : < Width, Pos, Length, Side > \quad (2)$$

(그림 3)의 경우, 상단면과 좌측면에서는 선분 에지가 존재하지 않으나, 하단면에서는 2개의 상향 수직 선분 에지가, 우측면에서는 3개의 좌향 수평 선분 에지가 존재한다. 하단면에서 선분 에지 검출 과정을 살펴보면, 흑백, 백흑 변화를 조사하여 wbA, bwA 쌍을 발견하고 이 구간을 첫 번째 수직 선분 에지의 폭으로 결정한다. [wbA, bwA] 구간의 각 점에서 시작하여 수직 방향으로 블록 내부의 연속된 검은 화소의 개수를 세면 BR(wbA)=5, BR(wbA+1)=9, BR(wbA+2)=3이며, 그 중 최댓값인 9가 선분의 길이이고 선분의 시점은 (wbA+1)의 위치, 즉 linePos_A가 된다. 따라서, 검출된 선분은 <3, 3, 9, bottom>의 속성을 갖는 상향 수직 선분이 된다. 하단면에서 검출된 두 번째 수직 선분(시점이 linePos_B인 선분)은 길이(2)가 블록 높이의 50%보다 짧아서 삭제된다. 우측면에서는 시점이 linePos_2인 수평 선분이 길이가 짧아서 삭제되고, 시점이 linePos_1인 선분과 linePos_3인 두 선분만이 유효한 수평 선분으로 검출되며, 이들의 속성은 각각 <3, 3, 9, right>, <1, 9, 9, right>이 된다.

(그림 3) 선분 에지와 꼭짓점 추출 과정



(Figure 3) Extraction of line segments and corner points

3.4 꼭짓점 검출 및 유형별 분류

에지 블록이 선분 에지를 포함하더라도 항상 꼭짓점을 포함하는 것은 아니다. (그림 1)의 "Non-feature" 블록에서와 같이 블록을 관통하는 직선만을 포함하거나, 블록의 한 면에서만 선분 에지가 나타나는 경우 서식을 구성하는 주요 특징점이 존재하지 않기 때문이다. 따라서 블록의 한 쪽 경계면에서만 선분 에지가 존재하는 경우와 서로 마주보는 두 경계면에서만 선분 에지가 존재하는 경우에는 제거하고, 나머지 블록들에 대해서만 특징 블록으로 고려한다.

특징 블록으로 추출된 블록에 대해서는 각 경계면에서 검출된 수평, 수직 에지 선분을 조합하여 꼭짓점의 위치를 결정하는데, 그 유형에 따라 꼭짓점 검출 과정에서 약간의 차이점이 있다. <표 2>의 유형 1~4의 경우, 블록의 두 경계면에서만 에지 선분이 검출이 되며, 이 때 검출되는 수평 선분을 L1:<L1.w, L1.p, L1.l, L1.s>으로, 수직 선분을 L2:<L2.w, L2.p, L2.l, L2.s>로 가정하면 꼭짓점의 좌표는 (L1.pos., L2.pos)로 구해진다. 유형 5~9의 경우에는 3~4개의 경계면에서 에지 선분이 존재하고 이들은 하나의 수평/수직 선분이 마주보는 두 경계면에서 검출되는 것이기 때문에 이들의 x 좌표 또는 y 좌표가 임계값 범위 내에서 일치할 경우에는 하나의 선분으로 통합된 후 꼭짓점의 좌표가 구해진다. 이 꼭짓점의 좌표로부터 선분의 시작점(2~4개)까지 끊김없이 검은 화소로 연결되어 있고, 각 선분의 길이가 임계값 이상일 경우 꼭짓점으로 판정한다.

<표 2> 꼭짓점 유형 분류

Type	Combinations of block sides
1	bottom ∧ right
2	bottom ∧ left
3	top ∧ right
4	top ∧ left
5	top ∧ bottom ∧ right
6	top ∧ bottom ∧ left
7	top ∧ left ∧ right
8	bottom ∧ left ∧ right
9	top ∧ bottom ∧ left ∧ right

<Table 2> Type classification of corner points based on the combination of line segments

이상의 과정에서 검출된 꼭짓점은 영상 위를 한 화소씩 이동하는 윈도우에 의해 생성된 블록으로부터 구해지기 때문에 동일한 꼭짓점이 여

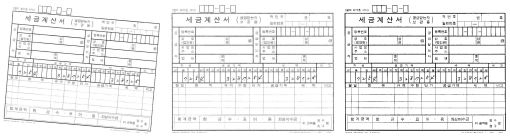
러 블록에서 수차례 검출된다. 따라서 전체 영상을 처리한 후 꼭짓점으로 검출된 횟수가 기준치 이상인 꼭짓점만을 최종 꼭짓점으로 선택한다. (그림 3)에서는 하단면의 linePos_A, 우측면의 linePos_1, linePos_3을 조합하여 꼭짓점 좌표 C1(linePos_A, linePos_1), C2(linePos_A, linePos_3)를 생성한다. 이 중, 꼭짓점 C2는 수직 길이가 2로 짧아서 무시되고, 꼭짓점 C1만이 꼭짓점 후보가 된다. 꼭짓점 C1을 기준으로 하단면의 linePos_A까지, 동시에 우측면의 linePos_1까지 끊김 없이 연결되었다면, 이 점이 유형 1의 꼭짓점(C1)이 된다.

4. 실험결과

제안한 방법을 C 언어로 구현하여 HP3020 복합기용 스캐너를 통하여 100~150dpi의 회색영상으로 입력 받은 세금계산서(150dpi, 960X720), 거래명세표(150dpi, 700X1012), 표를 포함하고 있는 일반문서(100dpi, 1248X488) 영상에 적용하였다.

(그림 4(a))은 일반적으로 통용되는 세금계산서 문서로 기울기 보정을 위하여 의도적으로 기울여서 입력받은 영상이다. 영상에 대한 기울기를 계산하고 이를 보정한 결과와 이진화한 결과 영상은 각각(그림 4(b), (c))와 같다.

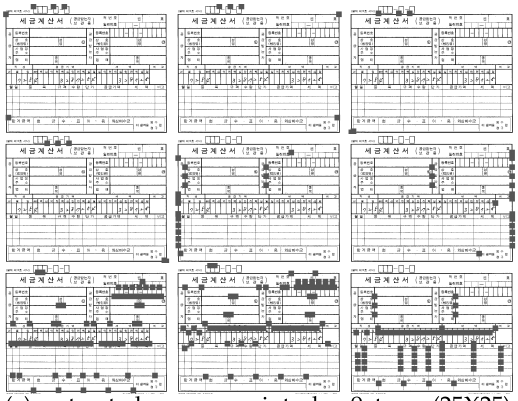
(그림 4) 세금 계산서 영상에 대한 전처리 결과



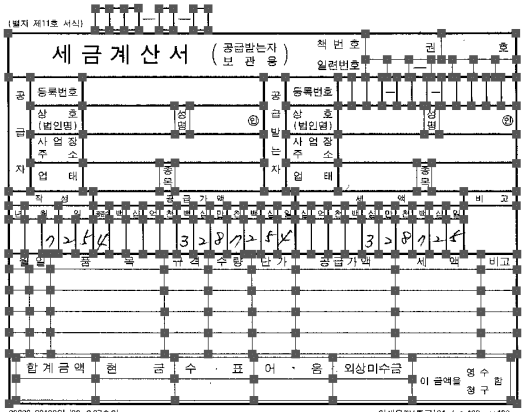
(a) Original (b) Skew-corrected (c) Binarization
(Figure 4) Preprocessing Results for Tax form image

(그림 5(a))는 입력 영상에 대한 9가지 유형의 꼭짓점을 추출한 결과이고, (그림 5(b), (c))는 마스크의 크기를 다르게 하여 적용하였을 때 추출된 꼭짓점을 모아서 한꺼번에 보인 결과이다. (그림 5(b))의 하단부(합계금액~청구함을 기록한 내부 표)를 보면 바깥쪽 표 안에 작은 표가 포함되어 이중 실선의 형태를 띠고 있고, 이러한 경우에도 정확하게 꼭짓점을 추출하고 있다.

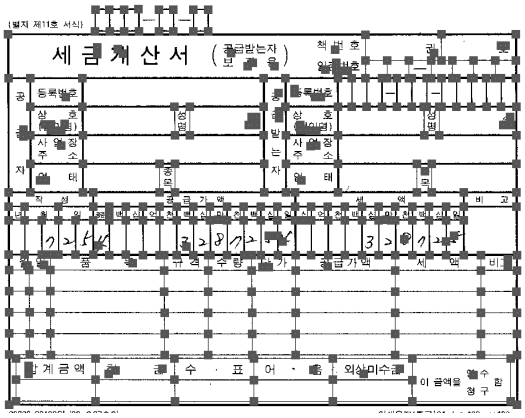
(그림 5) 특징점 추출 결과(세금계산서 영상)



(a) extracted corner points by 9 types(25X25)



(b) in the case of block size : :25X25



(c) in the case of block size ::15X15
(Figure 5) Extracted corner points(tax image)

이는 마스크가 영상 위를 한 화소씩 이동해나가면서 블록을 추출하고 꼭짓점을 검출하는데

마스크의 위치에 따라 여러 개의 꼭짓점을 포함하더라도 (그림 3)의 과정에 따라 선분의 길이가 임계값 이상인 모든 꼭짓점을 검출해내기 때문이다. (그림 5(c))의 경우 마스크의 크기를 15X15로 설정하여 실험한 결과로써, 수평/수직 선분으로 구성된 표 위에 존재하는 정확한 꼭짓점 외에 추가로 인쇄/기입된 문자 상에서도 꼭짓점들이 잘못 검출되고 있다. 이는 마스크의 크기가 너무 작을 경우 나타날 수 있는 문제로, 입력 영상의 크기에 따라 적절한 크기의 마스크를 설정하는 것이 필요하다는 것을 알 수 있다.

<표 3>에서는 몇 개의 세금 계산서 영상에 대해 적용한 결과를 보이고 있다. 여기에서 G.T.(Ground Truth)는 실제 꼭짓점의 개수를 의미하며, 굵게 표시한 부분은 검출 오류의 발생을 의미한다.

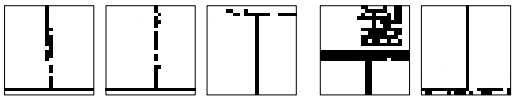
<표 3> 5개 영상에 대한 유형별 특징점 개수

Type / Images	1	2	3	4	5	6	7	8	9	total
G.T.	5	5	5	5	18	18	72	72	70	270
Image 1	5	5	5	5	18	18	72	72	70	270
Image 2	5	5	5	5	18	18	72	72	70	270
Image 3	5	5	5	5	18	18	**71**	72	**71**	270
Image 4	5	5	5	**6**	18	18	70	72	70	269
Image 5	5	5	5	5	18	18	**71**	**71**	70	268

<Table 3> Number of feature points extracted by 9 types for 5 sample images of tax form

총 1350개의 꼭짓점(5*270) 중에서 5개의 검출 오류가 발생하고 있어 99.7%의 검출 율을 보이고 있는데, 이중 3개는 미검출(detection miss) 오류이며, 2개는 오검출(false detection) 오류이다. 오류 위치에서의 부분 영상을 확대하여 보이면 (그림 6)과 같다. (그림 6(a))의 경우 실제로는 "ㅗ" 또는 "ㅜ" 유형의 꼭짓점을 포함하고 있으나 수평 또는 수직 선분 중간에서 연속 2 화소 이상의 끊김이 존재하여 임계값 이상의 긴 직선 선분을 검출하지 못하였다. (그림 6(b))는 첫 번째 영상은 "ㅜ" 유형을 "ㅏ" 유형으로, 두 번째 영상은 "ㅗ" 유형을 "ㄴ" 유형으로 잘못 검출한 결과이다. 이 중 첫 번째 그림은 서식에 인쇄된 문자와의 간섭 현상 때문인데, 선분과 문자와의 거리가 1 화소 이하이면서 문자가 끊김 없이 임계값 이상의 길이로 수평 또는 수직 선분을 포함하는 경우에 해당한다. 그 외의 경우 즉, 선분과 문자와의 거리가 2화소 이상 떨어져 있거나, 문자가 포함하는 선분이 사선 형태이거나 불연속적인 수평/수직 선분을 포함하는 경우에는 정확한 꼭짓점을 검출하고 있다.

(그림 6) 미검출/오검출 사례

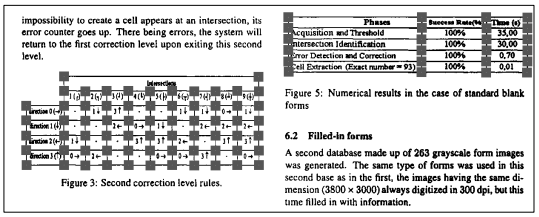


(a) detection miss (b) fault detection

(Figure 6) Examples of detection Miss/Falut

다양한 서식에 적용 가능성을 보기 위하여 (그림 7)은 임의의 논문지에서 발췌한 일반 문서를 스캔 받아 적용한 결과를 보이고 있는데, 표 내부에 기입된 텍스트의 간섭 없이 꼭짓점들을 정확하게 검출하고 있다.

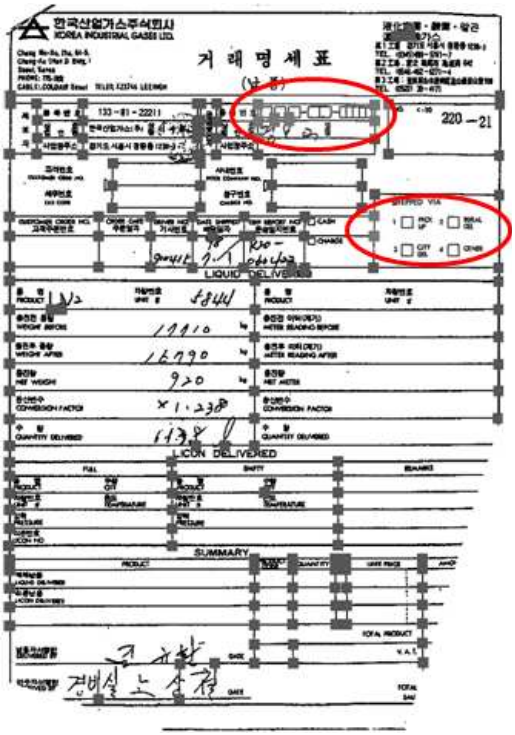
(그림 7) 특징점 추출 결과(일반 문서 영상)



(Figure 7) Extracted Feature points(ordinary document containing tables)

(그림 8)에서는 거래명세표 영상에 적용한 결과를 보였다. 그림에서 빨간 색으로 표시된 타원 내부의 체크 박스에서는 특징점이 검출되지 못하고 있는데, 이는 작은 사각형의 크기가 마스크 크기의 50%보다 작아서 사각형의 수평/수직 선분이 마스크 내에 최대한 포함되더라도 그 길이($Length$)가 마스크 크기의 50%보다 커야한다는 식(1)의 조건을 만족하지 못하여 배제되기 때문이다.

(그림 8) 특징점 추출 결과(거래명세서 영상)



(Figure 8) Extracted Feature points(transaction receipt image)

5. 결론

본 논문에서는 서식 문서 영상에 대한 선분 에지 기반의 유형별 꼭짓점 추출 방법을 제안하였다. 먼저, 입력된 서식 영상을 이진화한 후, 문서의 기울기를 계산하여 필요시 기울어짐을 보정한다. 이후, 잡음 등의 원인으로 나타날 수 있는 작은 크기의 검은 영역을 고립점으로 판단하여 삭제한 후, 유형별 꼭짓점 추출을 수행하였다. 본 논문에서 정의한 꼭짓점들은 선분 에지들이 직교하는 교차점들로 9가지 유형으로 분류하였다. 제안한 방법을 몇 가지 형태의 서식 문서에 적용하여 우수한 꼭짓점 추출 결과를 보였다. 이러한 유형별 꼭짓점 추출 결과는 서식 문서상의 대부분의 꼭짓점들이 대칭 형태로 존재한다는 사실을 고려할 때, 그들의 위치 관계를 활용하여 서식의 구조 분석에 매우 유용하게 활용될 수 있다.

향후에는, 서식 내에 마스크의 크기보다 작은 크기의 사각형에 대해서도 꼭짓점을 추출하기 위한 방안 마련이 필요하며, 처리 시간 단축을 위하여 현재 영상을 탐색하는 마스크의 이동 거리가 한 픽셀 단위인 점을 보완할 필요가 있다. 또한, 테이블의 최외각 꼭짓점이 직선 형태가 아닌 라운드 형태의 테두리를 가지는 테이블에 대한 처리, 상하 선분만 존재하고 좌우 선분이 존재하지 않는 형태의 테이블 처리 방식에 대한 연구가 추가적으로 필요하다.

References

[1] S. R. Hong, “The contrast between traditional printed text and hypertext reading comprehension”, Journal of Digital Contents Society, vol. 10, no. 4, pp. 537-542, 2009.

[2] R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Automatic processing of handwritten bank cheque images: a survey," Int. Journal on Document Analysis and Recognition, vol. 15, pp. 250-292, Jul., 2011.

[3] J. Chen and D. Lopresti, “Model-based ruling line detection in noisy handwritten documents”, Pattern Recognition Letters, vol. 35 pp. 34-45, 2014.

[4] S. Mandal, S. P. Chowdhury, and A. K. Das, "Fully automated identification and segmentation of form document," Computer Vision and Graphics, vol. 12, pp. 953-961, 2006.

[5] R. Palacios and A. Gupta, “A system for processing handwritten bank checks automatically”, Image and Vision Computing, vol. 26, no. 10, pp. 1297-1313, 2008.

[6] H. Nielson · W. Barrett, “Consensus-based table form recognition of low-quality historical documents“, International Journal of Document Analysis, vol. 8, no. 2, pp. 183-200, 2006.

[7] A. Amano, N. Asada, M. Mukunoki and M. Aoyama, “Table form document analysis based on the document structure grammar”, International Journal of Document Analysis, vol. 8, no. 2, pp. 201-213, 2006.

[8] T.Watanabe, Q. Luo, and N. Sugie, "Layout recogniti

on of multi-kinds of table form documents," IEEE Transactions on Pattern Analysis and Machine Intel ligence, vol. 27, pp.432-445, 2005.

[9] S. Taylor, R. Fritzson, and J. Pastor, "Extraction of data from preprinted forms," Machine Vision and A pplications, vol. 5, no. 3, pp.211-222, 1992.

[10] L. A. P.Neves and J. Facon, "Methodology of autom atic extraction of table-form cells," Brazilian Sympo sium on Computer Graphics and Image Processing (SIGGAPHI2000), pp.15-21, 2000.

[11] J. H. Ahn, "A simulation study on the fast gradient-based peak searching method", Journal of Digital Co ntents Society vol. 11, no. 1, pp. 39-45, Mar. 2010.

[12] J. Y. Jung, and M. Km, "Fast skew detection of document images by extraction of center points bet ween blank lines", Journal of KISS(B), vol.26, no.11, pp.1342-1349, Nov. 1999.

정 재 영

1989년 : 성균관대학교 (공학사)
1993년 : 성균관대학교 (공학석사)
1997년 : 성균관대학교 (공학박사-인공지능)
1997년～현 재: 동양대학교 컴퓨터정보전학과 교수
관심분야 : 인공지능(Artificial Intelligence), 패턴인식(Pattern Recognition), 움직임 추적(Motion Tracking) 등