2002년도 대한전기학회 하계학술대회 논문집 2002.7.10-12

레이블링기법을 이용한 문자 추출과 인식에 관한 연구

원혜경*, 김 용*, 이규훈**, 조규만***, 이은영****
동국대학교*, 스마트비젼텍**, 청주기능대학***, 대림대학****

A Study on the Character Extraction and Recognition using Labeling Method

Hye-Kyung Won*, Yong Kim*, Kyu-Hun Lee**, Kyu-Man Cho***, Eun-Yung Lee****
Dongguk University*, Smart Vision Tech**, Chongju Polytechnic College***, Daelim College****

Abstract - The process of character recognition goes through 5 steps; image acquisition, character region extraction, preprocessing, character region segmentation, character recognition. Therefore the final recognition rate of character recognition is directly affected by the performance of each step.
This paper is a leading research for object recognition using image processing algorithm which is one of the field of study in computer vision. And this paper will suggest an algotithm to extract the portion of number chain, which is part of the research embodying a system to perceive the date of manufacture and the name of the producer on the wrapping of groceries. In addition, this can extract the number chain comparatively accurate without using many complex algorithm by diving and extracting the moving number region at the same time.

1. 서 론

컴퓨터가 개발된 이후로 이론에만 그쳤던 많은 기술들이 이를 기반으로 현실에 응용되고 발전될 수 있었다. 그 중에서도 문자인식분야는 컴퓨터기술과 관련하여 다양한 응용분야의 도출 및 관련이론의 발전을 이루어 왔다. 또한, 여러 분야와 연계되어 인간의 생활을 풍요롭게 만드는 기술로의 가능성이 입증되어 왔고, 점점 지능화되어 가는 제품들에 필요한 기술로 인식되고 있다. 문자인식분야는 다른 응용분야에 비하여 상당히 오랜 역사를 가지고 있다. 특히, 오프라인 문자인식은 컴퓨터와 워드프로세서의 광범위한 보급으로 다양한 문서출력을 만들어 냈고, 사무자동화와 관련하여 그 응용분야를 무수히 창출할 수 있는 잠재성을 가지고 있으므로, 이미 제한된 분야에서 문자인식이 응용되고 있으며, 상용화된 제품이 나오기도 하였다. 그 결과 아직 기대이하의 인식률과 가격으로 널리 보급되지는 못하고 있는 실정이지만 사무자동화 뿐만 아니라 공장자동화에까지 다양한 시도가 이루어지고 있다. 또한 앞으로 인공지능분야와 관련하여 중요한 비중을 차지하게 될 분야이다.
최근 문자인식은 식료품 공장라인에서 식료품의 제조년, 월, 일등의 일련의 숫자를 추출, 검출하는 분야에서도 시도되고 있다.
기존의 문자영역 분리 방법에는 투영을 이용하여 분리하는 방법과 검사물의 사전지식을 이용하는 방법 그리고 두 가지 방법을 혼합하여 분리하는 방법등이 있고 그 외 다른 여러 가지 방법들이 있다. 이러한 방법들은 잡영에 민감하여 분리성공률이 낮고 계산량이 많아 처리속도가 매우 느리다. 상기 알고리즘들은 문자분리에 실패하였을 경우 동일 영상에 대하여 문턱값(threshold)을 반복하여 재조정하고 최악의 경우에는 여러 차례에 걸쳐 처리되는 경우도 있어 시스템 전체의 성능저하는 물론 비용증가의 커다란 원인이 된다.[1-4]
본 논문에서는 상기와 같은 문제점을 고려하여 인쇄된 문자중 일련의 숫자부분을 찾기 위하여 이진화된 영상을 기반으로 레이블링하여 숫자영역을 분할하고 추출하며, 실패할 경우 다시 수행함으로써, 여러 단계를 수행하여 추출할 수 있었던 일련의 숫자들을 비교적 빠르고 정확하게 추출하는 알고리즘을 제시하고자 한다.

2. 본 론

2.1 레이블링 기법

컴퓨터 비전에서, 레이블링은 특정 이미지의 연결된 구성요소를 찾는데 사용되는 일반적인 방법이다. 레이블링은 연결되어 있는 화소에 같은 레이블을 붙이고 다른 성분에는 다른 레이블을 붙이는 처리기법으로서 이 과정을 통하여 물체들을 개개의 연결 성분으로 나눌 수 있고, 각 연결 성분의 특성을 조사할 수 있다. 레이블링기법의 전제조건은 이진화가 되어야 한다는 점이며, 레이블링에는 재귀적 알고리즘과 순차적 알고리즘, 두 가지 방법이 있다. 재귀적 알고리즘은 알고리즘은 간단하나, 수행 시간에 있어서 비효율적이고, 순차적 알고리즘은 수행 시간은 적게 걸리나, 알고리즘이 매우 복잡하다. 본 논문에서는 간단한 재귀적 알고리즘을 사용하여, 재귀적 알고리즘의 비효율적인 면을 해결하는 알고리즘을 제안하고자 한다.
재귀적 알고리즘은 다음과 같다.
1. 영상을 살펴 레이블이 붙지 않은 화소 P를 발견하고, 새로운 레이블을 붙인다.
2. 화소 P와 연결되어 있는 화소에 같은 레이블을 붙인다.
3. 2에서 레이블을 붙인 화소와 연결된 전체 화소에 같은 레이블을 붙인다.
4. 이와같은 순서로 레이블을 붙여야 할 화소가 없어질 때까지 계속한다. 이것으로 하나의 연결성분 전체에 같은 레이블을 붙일 수 있게 된다.
5. 다시 1로 돌아가 아직 레이블이 붙지 않은 화소를 발견한다면 새로운 레이블을 붙여 2에서 4까지의 처리를 행한다.
6. 영상전체의 조사가 끝나면 처리를 완료한다.
그림 1과 같이 이진화된 이미지에 순서적으로 레이블을 붙인다.(그림 2)

```
00100100
00110100
00011000
01000100
01000000
```

그림1. 이진화된 이미지

00100200
00330400
00055000
06000700
08000000

그림2. 레이블링된 이미지

2.2 연결성분 추출 알고리즘

연결성분 추출처리를 하는 기본적인 기법은 레이블링(Labeling)방법이다. 레이블링 방법을 사용하는 경우 대량의 기억영역이 필요하게 되고, 또한 레이블링된 연결성분을 추출하기 위해서는 연결성분의 수 만큼 전화상 영역을 반복해서 주사할 필요가 있어 많은 처리시간이 요구된다. 이러한 단점을 보완하고자 다음과 같은 알고리즘을 사용하였다.

즉 그림 2와 같이 레이블링된 영상은 다음과 같은 레이블관계를 갖는다.

배경 레이블은 0으로 이것은 절대 변하지 않는다. 그림 2의 두 번째 행에서 레이블 3은 레이블 1과, 레이블 4는 레이블 2와 상관지어진다. 첫 번째 행에서의 결합메모리는 assoc[0..8] = {0,1,2,3,4,5,6,7,8}로 초기화되어있으며, 두 번째 행의 레이블간의 관계에 의해 assoc[0..8] = {0,3,4,1,2,5,6,7,8,}이 된다.

세 번째 행에서 레이블 5는 먼저 레이블 3과, 다음으로 레이블 4와 관계한다. 이에 의해 assoc[0..8] = {0,4,5,1,2,3,6,7,8}이 되며, 네 번째 행에서는 레이블 5가 레이블 7과 관계하므로, assoc[0..8] = {0,4,7,1,2,3,6,5,8}이 된다. 마지막으로 다섯 번째 행에서 레이블 8은 레이블 6과 상관하며 assoc[0..8] = {0,4,7,1,2,3,8,5,6}가 되서 더 이상 레이블링을 하지 않는다. 메모리는 1→4→2→7→5→3→1과 6→8→6의 두가지 주기를 갖게 되며 다음과 같이 최종 레이블링된다.

00100100
00110100
00011000
02000100
02000000

그림3. 최종 레이블링된 이미지

2.3 포장재질위에 인쇄된 일련의 숫자추출

본 논문에서는 레이블링할 때, 포장재질 위에 인쇄된 일련 숫자에 대한 사전정보를 가지고 다음과 같은 몇 가지 알고리즘을 제시한다.

먼저 레이블링의 재귀적 알고리즘의 단점인 긴 처리시간을 보완하기 위해 주위 4-이웃 알고리즘이나 주위 8-이웃 알고리즘을 사용하지 않고 현재 화소에서 같은 행과 열의 이전 화소와의 비교만으로 레이블을 붙인다.

재귀적 알고리즘을 사용할 경우, 처리한 화소에 대해 다시 레이블링하는 경우가 발생하여 속도저하를 발생시키고, 메모리상에 부하를 일으키기 때문에 이러한 경우를 대비하여 처리한 화소에 대해 플래그를 두어 처리한 화소에 대해서 여러번 수행하는 것을 방지할 수 있다.

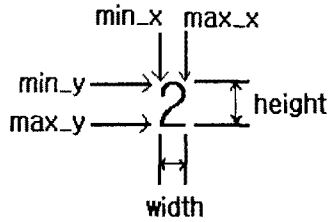

그림4. 인쇄된 숫자의 특성

그림 4와 같이 레이블링된 이미지에 대해 최소 X좌표, 최대 X좌표를 저장하여 넓이를 계산하도록 하였으며, 최소 Y좌표, 최대 Y좌표를 저장하여 높이를 계산한다. 또, 인쇄시 기본값으로 정해진 높이와 넓이의 정보로, 높이 대 넓이의 비율, 레이블링된 이미지의 픽셀의 개수의 정보를 가지고, 레이블링된 이미지에 대해 제약조건을 두어 일련의 숫자를 추출할 수 있다.

마지막으로 한번 더 레이블링을 수행함으로써 최종 레이블링된 이미지의 개수를 저장하여, 이 이미지개수로써 성공여부를 판단할 수 있다.

레이블링 수행후 일련의 숫자를 제외한 레이블링 이미지를 잡음으로 간주하고 잡음처리를 하여 최종 레이블링 이미지를 얻는다.

각 레이블링 된 이미지에 대해 X, Y좌표를 저장하였으므로, 일련의 숫자들의 X, Y좌표들은 어느 정도 일정한 간격을 유지하고 있다. 그러므로, 그 범위를 벗어난 레이블링 이미지를 잡음으로 처리할 수 있다. 따라서 그림 7과 같이 잡음을 제외한 최종 숫자들만 남게 된다. 각각의 숫자들은 레이블링되어 있으므로 따로 추출할 필요가 없으며, 레이블링할 때 X좌표와 Y좌표의 정보를 가지고 있으므로 각각에 대해서 간단하게 숫자들만 추출하고 검출해낼 수 있다. 기울어져 인쇄된 숫자에 대해서는 기울임 보정을 해준 후, 숫자를 검출한다.

그림 5는 그레이 칼라로 입력된 입력영상을 나타내며, 그림 6은 레이블링의 전제 조건인 이진화 영상을 나타낸다.

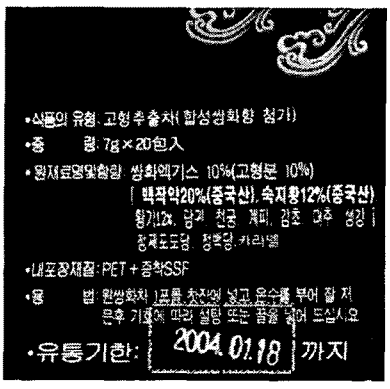

그림5. 그레이 이미지

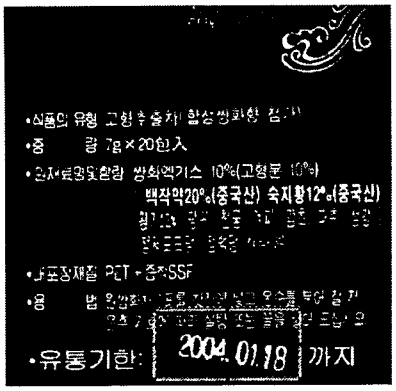

그림6. 이진화 이미지

레이블링기법을 이용하여 얻어진 이미지를 그림 7에 나타내었다.
각각의 일련의 숫자는 레이블링되어 있으므로 따로 추출할 필요가 없으며, 레이블링할 때 X좌표와 Y좌표의 정보를 가지고 있으므로, 각각에 대해서 간단하게 일련의 숫자영역만 그림 8과 같이 검출할 수 있다.

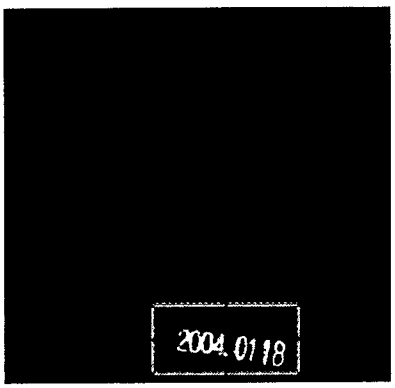

그림7. 레이블링 수행후 이미지

그림8. 검출 이미지

2.4 실험 환경 및 결과

실험에 이용된 영상 입력장치로는 유효화소 640×480(H×V)Resolution, Electric Shutter(1/30~1/10000), 무결점 CCD 카메라를 사용하였다. 본 알고리즘은 Pentium 1GMhz Main Board, 128MByte RAM, 20GByte HDD, AGP Video Card, WindowsNT환경의 시스템에서 일반 식료품의 포장재질을 캡쳐한 영상들을 가지고 Visual C++언어를 사용하여 수행하였다.
영상에서 분리성공은 문자모양 그대로 분리된 경우를 나타내고 부정확분리는 문자모양 그대로 분리가 되지 않은 경우를 나타낸다. 부정확분리의 경우 문자영역과 거의 흡사한 잡영을 함께 가지고 있는 경우나, 인쇄상태가 안좋아 문자의 일부만 남았을 경우 그 문자를 잡영으로 인식하는 경우였다.
실험 결과 포장재에 인쇄된 숫자영역을 찾아내는 데 0.2~0.3초의 시간이 걸렸으며, 인쇄위치가 심하게 왜곡된 경우를 제외한 대부분의 일련의 숫자에 대해서 98%의 높은 성공률을 보였으나, 일부 뭉어진 인쇄 문자의 경우, 숫자 "1"을 잡음으로 간주하거나, "2"를 "3"으로 인식하는 경우가 발생하였다.

3. 결 론

본 논문에서는 이진화된 포장재질 이미지의 전체 영상을 가지고 레이블링기법을 이용하여, 일련의 숫자를 추출하는 방법을 제시하였다. 본 논문의 장점은 여러 알고리즘을 사용하지 않고 단지 본 논문의 알고리즘만으로 일련의 숫자를 찾아 분할하고 추출하기까지 비교적 적은 수행시간을 필요로 하는 점이라 할 수 있다.
그러나 각 대상물의 포장재질에 따라 각각 다른 상태의 이진영상을 갖게 되므로, 대상물의 재질에 따른 조명을 잘 선택하여 보다 나은 이진화 영상을 갖을 수 있도록 하여 보완할 수 있을 것이다.

[참 고 문 헌]

[1] S. Liang, M. Ahmadi, and M.Shridhar, "Segmentation of Toughing Characters in Printed Document Recognition", Proc. of 2th ICDAR, pp.569-572, 1993.
[2] G. L. Martin and J. A. Pittman, "Recognizing Hand-Printed Letters and Digits Using Back-propagation Learning", Neural Computation, Vol.3, pp.258-256, 1991.
[3] 이영태, 최영우, "인쇄된 저화질 숫자 인식 연구', 1997년 숙명여자대학교 자연과학논문집 제8호 129-133
[4]이도엽, 김형재, 배익성, 이철희, 차의영, "변형된 Run Length Coding 기법을 이용한 이치화된 자동차 번호판 영상에서의 문자 분리", 한국멀티미디어학회, '98 봄 학술발표논문집
[5] 제성관, 박재현, 차의영 "레이블링기법을 이용한 차량 일련번호 추출", 한국정보과학회, 추계 학술 발표 논문집, 2000.10.
[6] 김도현, 강민경, 차의영, "기울기 보정과 블록 분할 합병을 통한 문자 추출", 한국정보과학회, 가을학술발표논문집, 제28권 2호, pp.424-426, 2001년 10월
[7] 염재훈, 이문호, "C언어를 이용한 영상신호처리", 대영사, pp90-102