슬라브 제품 정보 인식을 위한 문자 분리 및 문자 인식 알고리즘 개발

Character Segmentation and Recognition Algorithm for Steel Manufacturing Process Automation

최성후, 윤종필, 박영수, 박지훈, 구근휘, 김상우
SungHoo Choi, Jong Pil Yun, YoungSu Park, JeeHoon Park, KeunHwi Koo, Sang Woo Kim

Abstract – This paper describes about the printed character segmentation and recognition system for slabs in steel manufacturing process. To increase the recognition rate, it is important to improve success rate of character segmentation. Since Slabs' front area surface are not uniform and surface temperature is very high, marked characters not only undergo damages but also have much noise. On the other hand, since almost marked characters are very thick and the space between characters is only about 10 ~ 15 mm, there are many touching characters.

Therefore appropriate character image preprocessing and segmentation algorithm is needed. In this paper we propose a multi-local thresholding method for damaged character restoration, a modified touching character segmentation algorithm for marked characters. Finally a effective Multi-Class SVM is used to recognize segmented characters.

Key Words : Character Segmentation, Character Recognition, SVM, Foreground Analysis, Background Analysis

1. 서론

제철소의 연주 및 분괴 공정 등에서 생산된 슬라브(slab)에는 서로의 구분을 위하여 해당 슬라브에 대한 정보가 기재되어 있으며 호스트 컴퓨터에서 보내 온 슬라브의 번호와 현재 들어온 슬라브에 기재되어 있는 번호의 일치 여부를 검사하기 위해 자동 인식 시스템이 설치되어 있다. 기존의 자동 인식 시스템은 슬라브의 사전 정보를 이용하여 인식 실패를 체크할 수 있으며, 이때는 피드백 알고리즘으로 슬라브에 표시되어 있는 번호를 재분리하고 검증하는 방식이다[2]. 사전 정보를 이용함으로써 사전 정보 오류를 예방하는 사전 정보 검증 부분이 포함되는 등 오히려 인식 알고리즘이 복잡해지는 경향이 있다. 본 논문에서는 사전 정보를 이용하지 않고도 효과적인 성능을 가지는 인식 알고리즘을 제안한다.

제철소의 열악한 환경 즉 조명의 변화, 장비의 이동 등에 의해 슬라브에 마킹된 물류 번호는 지속적인 훼손을 받게 된다. 특히 슬라브 자체의 높은 온도에 의해 번호의 상하 밝기 차가 발생하여 이진 영상 변화 시 번호 일부분이 소실되는 현상이 일어나기도 한다. 이러한 현상으로 인해 전처리 알고리즘에서 개별 번호를 분리하는 데 실패하거나, 또는 분리에 성공하더라도 훼손된 번호로 인해 최종 인식률이 현저히 떨어진다. 이러한 현상을 해결하기 위해 본 논문에서는 개별 번호 추출 단계에서 Background 분석법을 이용하여 환경 노이즈에 의해 겹쳐진 번호를 효과적으로 분리하는 방법을 이용하였다[4]. 또한 번호의 부분별 밝기 변화에 따른 번호 일부분이 소실 현상을 개선하기 위해 Multi-Local Threshold 방법을 이용하였다. 최종 인식에는 여러 응용분야에서 강력한 분류 능력을 보이는 SVM(Support Vector Machine)을 적용하였다[5].

본 논문의 구성을 보면, 제 2장에서는 각 구성 요소별 주요 알고리즘을 설명한다. 제 3장에서는 실험 결과를, 마지막으로 제 4장에서는 결론으로 본 논문을 마무리한다.

2. 전처리 글자 분리 알고리즘

2.1 전처리 알고리즘

슬라브의 전면 부에 마킹된 물류 번호는 슬라브의 온도가 매우 높을 때 마킹 머신에 의해 마킹되며, 이 때 마킹된 번호는 시간이 지나면서 서서히 훼손된다. 슬라브 전면 부의 거친 표면에 의해 글자 외에 원치 않는 수많은 노이즈들이 생기는 경우가 있으며, 혹은 접촉이나 열에 의해 또는 마킹 머신의 기계적인 한계로 인하여 문자의 일정 부분이 소실되는 경우가 매우 많이 발생한다. 이것은 곧바로 명도 영상에서 글자 자체의 밝기 변화를 일으켜, Otsu's Method를 이용한 단순 이진 영상 변환에서 훼손된 부분이 사라지게 하는 현상을 초래한다. 이러한 문제는 문자 분리를 완벽하게 성공하였다 하더라도, 최종 인식 단계에서 실패할 확률이 높다. 따라서 이러한 문제점을 해결하기 위하여 Multi-Local Threshold(이하:MLT) 방법을 제안한다. 알고리즘을 정리하면 다음과 같다.

(1) 명도 영상의 영역을 일정한 크기를 가지는 영역으로 분할하여 각 영역의 Local Threshold 값을 구한다.

저자 소개
* 최성후 : 포항공대 전자과 통합박사과정
** 윤종필 : 포항공대 전자과 통합박사과정
*** 박영수 : 포항공대 전자과 박사과정
**** 박지훈 : 포항공대 전자과 석사과정
***** 구근휘 : 포항공대 전자과 석사과정
******김상우 : 포항공대 전자과 교수

(2) 앞서 구한 각 영역의 Threshold 값을 이용하여, 이진 영상으로 변환한다.

(3) 나누어진 각 영역을 본래의 위치에 맞도록 병합하여 이진 문자열 이미지 데이터를 얻는다.

본 논문에서는 이 방법을 문자를 분리하기 전에 문자열 데이터에 대하여 적용하였고, 문자 분리가 끝난 후 분리된 개별 문자에 대해서도 적용하였다. 문자열 데이터에 적용할 때는, 영역을 작게 잡을수록 원치 않는 노이즈까지 되살리는 경우가 많으므로 적당한 크기로 영역을 나누어 주어야 한다. 실험 데이터의 분석을 통해 개별 문자에서는 좌-우의 밝기 변화보다는 상-하의 밝기 변화가 심했기 때문에, 이를 고려하여 상-하로만 영역을 나누어 Multi-threshold 방법을 적용하였다.

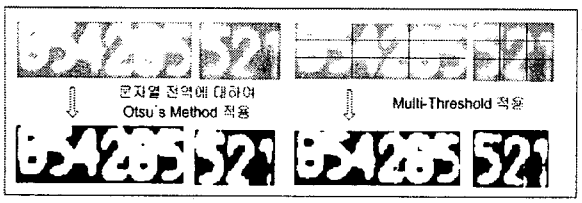

그림 1 Otsu's Method를 이용한 이진 영상 변환과 Multi-Threshold를 이용한 이진 영상 변환 차이

2.2 문자 분리 알고리즘

가열로에 장입되는 슬라브의 캡쳐 영상에서 취득된 문자열 데이터는 그림 1과 같다. 마킹된 인쇄체는 6번째 숫자와 7번째 숫자사이의 공백이 크므로, 본 논문에서는 그 점을 이용하여 1~6번째 문자 영역, 7~9번째 문자 영역으로 나누어 각각의 문자 영역에서 개별 문자를 분리한다. MLT 방법을 통해 얻어진 이진 영상과 명도 영상에서 Canny 경계 검출 방법을 통해 얻어진 경계선(Edge) 영상에서 흰 픽셀의 수직 투영 프로파일을 구하며, 특히 이진 영상에서는 검은 픽셀의 수직 투영 프로파일을 더 구하는데, 이것은 각 문자열 이미지 데이터의 윗부분에서 아래를 내려 볼 때 흰 픽셀이 발견되면 그 픽셀부터 이미지 데이터의 아래 부분까지 모두 흰 픽셀로 만들었을 때 얻어지는 영상에서 세로 방향으로 검은 픽셀의 개수를 더함으로서 구할 수 있다(그림 2).

슬라브에 마킹되는 인쇄체는 열에 의해 소실되는 것을 방지하기 위해 두껍게 마킹되며, 슬라브의 폭이 가장 좁을 때도 마킹이 가능하도록 하기 위해 글자 간격이 매우 좁다. 따라서 문자들 간의 겹침 현상이 빈번하게 발생하게 되고, 만일 겹침 현상이 발생하지 않더라도 위-아래를 잇는 일직선으로 글자 구분이 불가능한 경우가 생긴다. 이러한 문제를 해결하기 위해 앞서 구한 두개의 수직 투영 프로파일로부터 각 글자 사이의 최적의 선형 경계(Linear Boundary) 혹은 비선형 경계 후보 영역을 찾아내고, 글자를 제외한 배경의 형태학적인 해석과 관련된 Background 분석법을 이용하여 최적의 비선형 경계를 찾아낸다[4] 본 논문에서는 글자 자체의 형태학적인 분석과 관련된 Foreground 분석법은 사용하지 않았으며, 참고 문헌에서 사용한 알고리즘인 Foreground 및 Background 분석법을 수정하여 사용하였다. 분리 알고리즘을 정리하면 다음과 같다.

(1) 문자열 이미지 데이터의 이진 영상, 경계 영상 획득

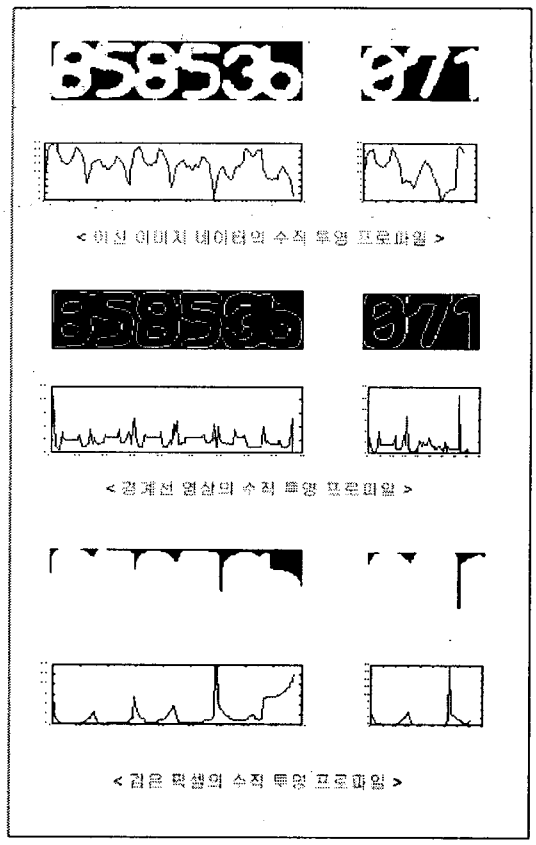

그림 2 문자열 이미지 데이터의 수직 투영 프로파일

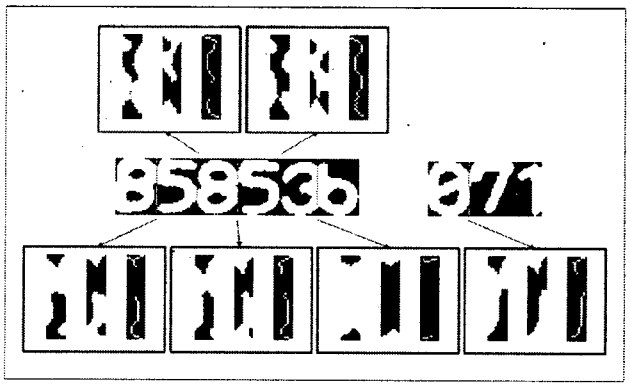

그림 3 선형 및 비선형 개별 문자 분리 과정

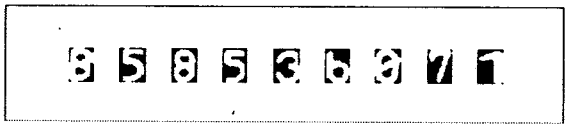

그림 4 분리된 개별 문자

(2) 이진영상으로부터 흰 픽셀의 수직 투영 프로파일과 검은 픽셀의 수직 투영 프로파일을 얻고, 경계 영상으로부터 흰 픽셀의 수직 투영 파일을 얻는다.

(3) 실험적으로 얻어진 글자의 최소 폭과 최대 폭을 고려하여 각 글자 사이의 경계 후보 영역을 얻고, 이 영역에서의 경계 영상의 수직 투영 프로파일, 이진 영상의 수직 투영 프

로파일을 차례로 분석하여 프로파일 값이 '0'이 되는 부분이 있는지 검사한다.

(3).a. '0'이 검출된 경우 : 글자 사이에 간격이 존재하여 선형 경계로서 나눌 수 있는 지점으로 판단하고, 선형 경계로 두 글자 사이를 나눈다.

(3).b '0'이 검출되지 않은 경우 : '0' 이 검출되지 않으면 글자 사이에 겹침이 발생하여 선형 경계로 나눌 수 없는 지점인 것으로 판단하고, 비선형 경계를 추출하기 위하여 이진 영상의 검은 픽셀 수직 투영 프로파일에서 극대점을 추출한 후, 극대점이 위치하는 지점을 중심으로 실험적으로 찾아낸 좌-우 마진을 두어, 비선형 경계 추출을 위한 후보 영역을 찾아낸다. 비선형 경계 후보 영역에서 후보 영역을 반전시키고 배경의 Thinning Line을 얻어내기 위해 형태학적 변환을 한다[3]. Thinning Line의 각 포인트를 비선형 경계선으로 결정하고 글자의 겹침이 있는 부분은 위·아래에 존재하는 Thinning Line의 끝부분을 서로 일직선으로 연결해줌으로서 겹침이 발생한 두 글자 사이의 최적의 경로를 찾아낸다.

(4) (1)-(3)을 (글자 수 - 1)번 만큼 반복한다.

(5) 찾아낸 (글자 수 - 1)개의 경계를 명도 영상에 적용하여 명도를 가지는 문자를 추출하고 개별 문자 영역을 일정등분으로 나누어 MLT를 적용하여 이진 영상으로 변환한다. 마지막으로 얻어진 개별 문자의 이진 영상은 인식기의 입력 형식에 맞도록 정규화 된다.

2.3 인식 알고리즘

Support Vector Machine(이하:SVM)은 2개의 클래스만을 구분하는 이진 분류기이며, 클래스 사이의 경계면은 Margin이 최대로 되도록 설정하여 경험적인 위험을 최소화하는데 기초를 둔 인식 알고리즘과는 달리 구조적 위험을 최소화 하는 것에 기반을 두고 있다[1,5]. 다중 클래스를 구분해야 하는 경우에 대해서는 이진 분류기인 SVM의 구조를 확장한 Multi-Class SVM을 이용하여야 하며, 여기서는 기본적이면서도 가장 성능이 가장 뛰어난 One-to-One SVM을 이용하였다[1].

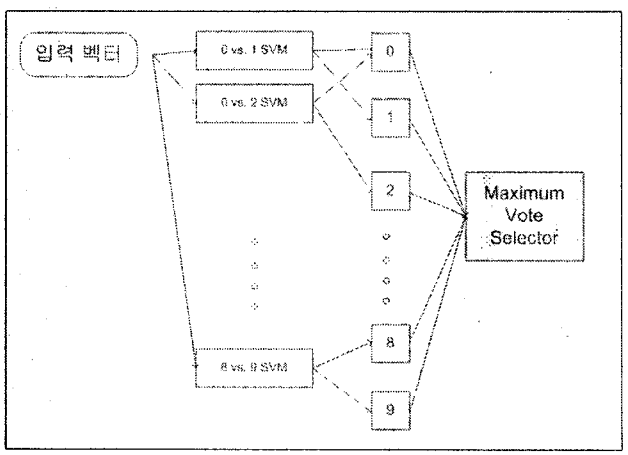

그림 5 One-to-One SVM

3. 실험 결과

Multi-Class SVM를 학습시키는데 있어서, 각 숫자 당 1000개씩의 트레이닝 데이터를 사용하였으며, 이 트레이닝 데이터를 포함하지 않은 슬라브의 문자열 이미지 데이터에 대해 테스트를 하였다. 표 1 과 표 2 에 개별 숫자에 대한 인식률과 슬라브 인식 성공률을 조사하였다.

숫자	0	1	2	3	4	5	6	7	8	9
인식률	97.4	97.2	99.7	99.2	99.5	98.5	94.7	98.9	97.1	98.0

표 1 개별 숫자별 인식률 (단위 : [%])

분류	인식률
모든 숫자 성공한 슬라브	89.4 %
1개의 숫자 실패한 슬라브	5.2 %

표 2 슬라브별 성공률

'6'을 제외한 나머지 숫자들이 모두 높은 인식률을 가짐을 확인하였고, 이것은 곧 겹침 현상이 발생한 숫자에 대해서 수정된 Background 분석법이 숫자를 분리함에 있어 큰 역할을 한 것으로 볼 수 있다. 각 숫자의 높은 인식률은 곧 슬라브 당 인식률에 영향을 미쳤으며, 사전정보를 활용하지 않더라도 순수 인식단계에서 거의 90%에 가까움을 보여주었다.

4. 결론

본 논문에서는 슬라브 번호 인식률 향상을 위해 전처리 과정에서 MLT를 사용하여 밝기 차이가 있는 문자 영역이나 개별 문자에 대해 단순 이진화시 소실되는 것을 방지하였고 문자 분리 시 겹쳐진 번호로 인한 개별 번호 추출 알고리즘의 오류를 개선하기 위해 Background 분석법을 이용하였다. 분류 성능이 우수한 One-to-One SVM을 사용하였으며, 실험 결과 기존의 피드백 알고리즘과는 달리 사전 정보를 이용하지 않고도 우수한 인식률을 나타냄을 보였다.

5. Acknowledgement

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음. (IITA-2006-C1090-0602-0013)

참 고 문 헌

[1] 이영교, 장유진, 김연탁, 김상우, "빌렛에서의 필기체 인식 시스템 개발", 2006 제어자동화시스템 심포지엄 CASS'2006, 1~3. June 2006, KINTEX, Korea.

[2] 홍기상, 장정훈, 양종렬, 김승진, 김태원, "슬라브 번호 인식 장치의 개발", 제어·자동화·시스템공학회지 6호, 제 2 권, pp. 63-76, 1996. 11.

[3] Rafael C. Gonzalez and Richard E. Woods, "Digital Image Processing, Second Edition", Prentice Hall

[4] Yi-Kai Chen and Jhing-Fa Wang, "Segmentation of Single- or Multiple-Touching Handwritten Numeral String Using Background and Foreground Analysis", IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 22, no. 11, pp. 1304-1317, 2000. 11.

[5] Nello Cristianini and John Shawe-Talyor, " An Introduction to Support Vector Machines", Cambridge University Press, 2000.