저해상도 인쇄체 한글 영상 인식을 위한 자소 분할 방법

이 성 훈[O], 조 규 태, 김 진 식, 김 진 형 (한국과학기술원 전산학과)
정 철 곤, 김 상 균, 문 영 수, 김 지 연 (삼성종합기술원 컴퓨팅랩)
{leesh[O],ktcho,jskim,jkim}@ai.kaist.ac.kr,{cheolkon.jung,skkim77,mys66,jiyeun.kim}@samsung.com

Grapheme Segmentation Method for Low Quality Printed Hangul Text Recognition

Lee Seong Hun[O], Cho Kyu Tae, Kim Jin Sik, Kim Jin Hyung,
Department of Computer Science, Korea Advanced Institute of Science and Technology
Cheolkon Jung, Sang-Kyun Kim, YoungSu Moon, Ji-Yeun Kim
Samsung Advanced Institute of Technology (SAIT)

요 약

본 논문에서는 저해상도 한글 영상을 자소 단위로 분리하는 방법을 제안한다. 비디오 자막이나 저해상도 스캔 영상의 경우 자소간 획이 접촉되거나 잡영이 많이 포함되어 기존의 자소 분할 방법으로는 한계가 있다. 한자 문자열을 문자 단위로 분할하는데 사용된 비선형 분할 경로 알고리즘을 한글 낱자 영상에 적용하여 자소 단위로 분할한다. 기존의 분할 경로 알고리즘을 한글 자소 분할에 효과적으로 적용하기 위해서 우세점 탐지 알고리즘을 이용하여 자소간 접촉점을 찾고 이를 바탕으로 생성된 분할 경로에 따라 여러 개의 자소 후보 영상이 생성된다. 자소 영상을 자소 인식기로 인식한 결과 높은 인식률을 보이는 것을 실험을 통하여 확인하였다.

1. 서 론

최근에 휴대폰 카메라를 비롯한 다양한 멀티 미디어 입력 장치의 개발로 인하여 기존의 스캐너와 같은 입력 방법에서 벗어나 다양한 정보 획득이 가능해졌다. 예를 들어, 수업 시간에 강의 내용을 휴대폰으로 찍어서 저장하거나 비디오 자막을 인식하여 내용을 요약을 하는 등 단순히 문서를 스캔 했을 때와는 다른 형태의 정보 습득 방법이 가능해졌다. 이때 입력된 한글 영상이 기존의 스캔을 하여 얻은 영상과는 달리 해상도가 낮은 경우가 많다. 영상의 해상도가 낮아짐에 따라 글자의 획이 접촉되거나 글자의 형태가 왜곡되는 현상이 발생한다. 하지만 기존의 인쇄체 한글 자소 인식 방법은 해상도가 낮은 낱자 영상에서 자소간 획을 정확히 분리하지 못하는 한계로 인해 오인식 되는 경우가 발생한다. 따라서 획이 붙어 있는 낱자 영상을 정확히 분리할 수 있는 방법이 필요하다.

본 연구에서는 저해상도 한글 영상을 인식하기 위한 자소 분할 방법을 제안한다. 제안한 방법에서는 한글 구조에 따라 6형식으로 분류된 낱자 영상을 각 형식별로 적합한 분할 기법을 이용하여 여러 개의 자소 후보로 분할한다. 여러 개의 분할된 자소 후보 영상은 자소 단위 인식기의 입력으로 사용된다.

본 논문의 구성은 다음과 같다. 2장은 기존의 한글 자소 분할 방법에 대해서 설명한다. 3장에서는 제안하는 시스템 구조를 설명한다. 4장은 자소 분할을 위한 비선형 분할 방법에 대해서 설명하고 이를 저해상도 영상에 적합하도록 개선한 방법을 설명한다. 5장에서는 실험 결과를 보인다.

2. 기존 한글 자소 분할 방법

한글 문자를 인식할 때 낱자 단위로 인식할 경우에는 인식할 대상 수가 많고 다양한 글꼴인 경우에 인식률이 낮아지는 한계가 있었다. 이를 해결하기 위해서 한글을 자소별로 인식하려는 방법이 많이 사용되었다. 한글을 자소별로 인식할 경우에는 자소의 수가 한정되어 있기 때문에 낱자 단위로 인식하는 것보다 인식 대상 수를 크게 줄여 인식 성능을 높일 수 있다. 자소 단위로 분할하는 방법에는 한글의 구조적 정보를 이용한 방법[1][2]이나 배경 세선화 방법[3] 등이 이용되었다.

한글 구조적 특성을 이용한 방법은 각 형식에 따라 각 자소의 상대적인 위치는 일정하다는 특징을 이용하여 자소의 대략적인 위치를 찾아서 해당 자소 영역을 분리하는 방법이다[1]. 하지만 이 방법은 해당 자소 외에 다른 자소 성분의 일부가 포함이 되어 커다란 잡영의 역할을 하여 인식하는데 방해를 준다. 이를 개선하기 위해서 자소 영역을 확대하여 인식하는 방법도 제시되었다[2].

배경 세선화 방법은 배경을 세선화하여 골격선을 추출한 후

문자의 접촉된 부분을 찾아내어 분할하는 방법인다[3]. 이 방법을 이용하면 자소간의 접촉된 부분도 분리할 수 있는 장점이 있다. 하지만 저해상도 영상과 같이 잡영이 심한 영상인 경우는 배경 골격선이 잡영에 영향을 받아서 잘못 추출되어 자소 영상을 제대로 분리할 수 없는 한계가 있다.

3. 시스템 개요

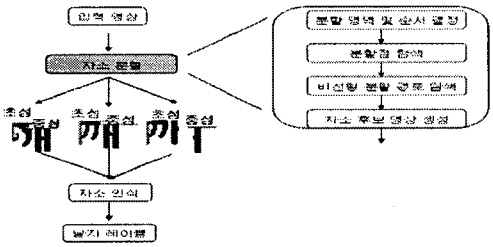

그림 1. 한글 인식 과정

한글 인식 과정의 전반적인 개요는 그림 1과 같다. 한글을 자소 단위로 인식하기 위해서 낱자 영상을 자소 영상으로 분할하는 과정이 필요하다. 자소 분할 과정은 세부적으로 4단계로 구분할 수 있는데 첫번째 단계에서는 한글의 형식에 따라서 분할 영역 및 분할 순서를 정한다. 두번째 단계에서는 분할 경로를 찾을 때 중요하게 이용되는 자소간 분할 점을 찾는다. 이러한 분할점은 외곽선의 곡률이 큰 특징을 이용하여 우세점 탐지 알고리즘[4]을 적용하여 찾을 수 있다. 세번째 단계에서는 많은 분할점 중에서 불필요한 분할점을 제거한 후에 개선된 비선형 분할 방법[5]을 적용하여 자소별로 분할한다. 획 접촉이 심한 글자인 경우는 하나의 비선형 분할 경로로 완벽하게 자소 영상을 분할 할 수 없다. 따라서 이를 여러 개의 자소 후보 영상으로 분할하여 그 중에 인식률이 가장 높은 영상을 해당 자소 영상으로 선택하고 그때의 인식 결과를 최종 레이블로 선택한다.

4. 자소 분할

자소 단위로 인식하기 위해서는 낱자 영상을 자소 영상으로 분할하는 과정이 필요하다. 낱자 영상이 오분할될 경우 다른 자소에 의한 잡영으로 오인식될 수 있으므로 자소간의 겹쳐진 부분이나 접촉된 부분에 대해서 정확히 분할할 수 있어야 한다. 이를 위해 자소 분할 과정에서는 여러 개의 분할 경로를 생성하여 그 분할 경로에 따라 자소 가능성이 높은 후보 영상들을 생성한다. 정분할된 영상이 오분할된 영상보다 높은 인식 확률 값을 가지는 특징을 이용하여 여러 개의 자소 후보 영상을 인식 단계에서 이용할 수 있다.

낱자 영상을 자소 영상으로 분할하기 위해 한글의 구조적인 정보를 이용하여 한글 형식에 따라 분할 종류 및 순서를 선택한다. 예를 들어 종성이 있는 낱자 인 경우는 수평 분할을 수행해서 종성을 먼저 분리하고 수직/수평 분할을 수행해서 초성과 중성을 분할한다. 또한 한글 형식별로 낱자 내 대략적인 분할 영역을 설정하여 그 영역 안에서 분할 경로가 생성되도록 하여 잘못된 경로에 의해 오분할 되는 것을 방지할 수 있다. 분할 순서 및 종류가 결정되면 낱자 영역을 자소 영역으로 분할하는 단계를 수행한다. 본 연구에서는 이를 비선형 분할 경로 알고리즘을 적용하여 자소간 분할 경계선을 구한다.

4.1 비선형 분할 경로

기존의 비선형 분할 경로 기법은 문자열을 글자 단위로 분할하기 위해서 Tseng에 의해서 고안된 방법으로 획의 접촉을 최소화하기 위해 분할 경로가 직선이 아닌 경계선이 되도록 하는 방법이다[5]. 이를 위해서 낱자를 다단계 그래프로 각 화소를 그래프 상의 노드로 표현하고 각 화소 사이의 경로를 노드의 전이로 표현한다. 백화소인 노드에 흑화소 노드보다 높은 점수를 배정하고 화소간 수평/수직 방향으로의 전이에 다각선 방향으로의 전이보다 높은 점수를 배정한다. 따라서 흑화소를 적게 지나갈수록, 대각선 방향으로 적게 전이될수록 높은 배정 점수를 받는다. 그래프 상의 모든 경로의 점수를 앞에서 정의한 화소 점수와 전이 점수를 이용해서 계산하고 이 점수가 최대가 되는 경로들을 동적 프로그램(dynamic programming)을 이용해서 선택한다.

예를 들어서, 그림 2의 첫번째 그림은 " 더" 원영상이고 두번째 그림은 비선형 분할 알고리즘을 통해 얻은 후보 경로이다. 여러 후보 경로 중에 겹쳐지거나 인접한 경로를 제거하면 네번째 그림과 같이 최종적으로 분할 경로를 구할 수 있다. 하지만 기존의 비선형 분할 방법은 획분할을 최소화하는데 주안점을 두기 때문에 네번째 그림과 같이 자소간 붙어 있는 획을 분할하지 않고 다른 획 중간을 통과하는 경우가 발생한다.

[원영상] [후보 경로] [겹쳐진 경로 제거] [인접 경로 제거]

그림 2. 비선형 분할 경로

4.2 한글에 적합한 개선된 비선형 분할 경로

비선형 분할 경로에 사용되는 노드와 전이에 점수를 부여하는 방식을 보완하여 한글 자소 분할에 적합하도록 분할 경로를 개선할 수 있다. 노드간의 전이를 고려할 때, 대각선 노드뿐만 아니라 대각선의 인접 이웃 노드까지 대각선 방향으로 함께 고려한다. 이처럼 대각선 경로 폭을 확대함으로써 자소간 겹침이 심한 경우에도 분할이 가능하다. 각 노드(화소) 중에서 자소간 붙어있는 획의 경계점이라고 생각되는 노드는 백화소나 흑화소보다 더 높은 점수를 부여해서 비선형 경로가 해당 노드를 지나가도록 유도한다. 이 방법을 통해서 기존에 경로가 획의 중간을 통과하는 오류를 감소시킬 수 있다.

붙어 있는 획의 경계점에서 외곽선의 기울기 변화가 큰 특징을 이용하여 경계점의 후보를 찾을 수 있다. 경계점을 찾는 방

법은 다음과 같다. 전처리 단계에서는 잡영으로 인해 글자의 외곽부분에 생긴 미세한 돌출 부분을 수직/수평 black run length를 이용하여 제거한다. 잡영이 제거된 영상으로부터 외곽선을 구하고 외곽선에 있는 모든 점들의 체인 코드를 구한다. 이 체인 코드에 우세점(dominant point) 탐지 알고리즘을 적용하여 방향각이 급격하게 변화된 점을 추출한다[4]. 많은 우세점 중에서 특정 범위 안에 있는 이웃 점들보다 방향각의 변화율이 낮은 점은 불필요한 점으로 간주하고 제거한다. 그러나 적용된 우세점 탐지 알고리즘은 자소의 경계점이 아닌 점들도 함께 우세점으로 추출되는 한계를 지니고 있다. 이를 한글 구조적 정보를 이용해서 우세점의 위치에 따라 노드에 점수를 차등적으로 부여하는 방식으로 이러한 문제를 줄일 수 있다.

그림 3은 개선된 비선형 분할 경로 알고리즘을 적용하여 자소간 분할 경로를 찾은 예이다. 비선형 분할 경로를 경계선으로 하여 자소 영상으로 분할할 수 있다.

그림 3. 개선된 비선형 분할 경로

5. 실 험

본 논문에서는 자소 분할에 대한 성능 평가를 위해 문자 인식기의 성능을 측정하였다. 실험에 쓰인 데이터로는 비디오 뉴스 자막과 저해상도로 스캔한 영상을 이용하였고 전체 25만개 정도로 영상의 크기나 글꼴이 다양하다. 인식기로는 낱자 영상을 인식하는 비분할 자소 인식기와 자소 영상을 인식하는 분할 자소 인식기를 이용하여 비교하였다[6]. 실험 성능의 정확도를 높이기 위해서 3-cross validation을 수행하였다.

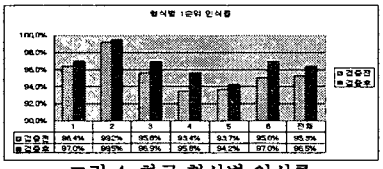

그림 4. 한글 형식별 인식률

테스트 영상에 대한 각 형식별 인식 결과를 그림4에 표시하였다. 왼쪽 막대 그래프는 낱자 영상을 대상으로 하는 비분할 자소 인식을 통해 얻은 인식률이다. 오른쪽은 분할된 자소 영상을 대상으로 하는 분할 자소 인식을 통한 검증으로 구한 인식률이다. 비분할 자소 인식을 수행하여 95.3%, 분할 자소 인식 방법을 이용한 검증을 통해서 96.5%의 성능을 얻은 것을 알 수 있다. 이는 낱자 영상을 대상으로 하는 비분할 자소 인식에 비해서 약 27.8%정도 오류가 감소한 것이다. 형식별로 살펴보면 자소의 구조가 복잡한 낱자일수록 자소 분할 후 재인식을 통해 오류가 많이 개선된 것을 알 수 있다.

그림 5는 제안한 자소 분할 알고리즘을 적용하여 분할에 성공한 예를 나타낸다. 자소의 글꼴이나 위치의 변이에도 자소 단위로 잘 분할하는 것을 알 수 있다. 또한 자소간 겹쳐지거나 접촉된 글자도 잘 분해하는 것을 알 수 있다.

그림 5. 정분할된 영상

그림 6는 자소 분할이 실패한 예이다. 획의 접촉이 심하거나 획의 소실로 자소 분할의 오류가 발생하였다. 이러한 자소 분할 오류로 인해 오인식이 발생하였다.

그림 6. 오분할된 영상

6. 결 론

본 논문에서는 저해상도 문자를 인식하기 위한 새로운 자소 분할 방법을 제안하였다. 문자열을 문자 단위로 분할하는데 효과적인 비선형 분할 경로 알고리즘을 한글 자소 분할에 맞게 개선하여 자소간 겹쳐지거나 접촉된 경우에도 잘 분할하는 것을 알 수 있었다. 특히 자소의 구조가 복잡한 낱자도 잘 분할을 하여 한글을 비교적 신뢰성 있게 인식할 수 있음을 실험을 통해서 확인하였다.

참고 문헌

[1] 권재욱, 조성배, 김진형, " 계층적 신경망을 이용한 다중 크기의 다중활자체 한글 문서 인식" , 한국정보과학회 논문지, 제 19권, 제 1호, pp.69-79, 1992

[2] 이판호, 장희돈, 남궁재찬, " 동적 자소 분할과 신경망을 이용한 인쇄체 한글 문자 인식에 관한 연구" , 한국통신학회논문지, 제 19권, 제 1호, pp. 2133-2145, 1994

[3] 이근수, " 퍼지 추론을 이용한 인쇄체 한글 인식" , 숭실대학교 박사 논문, 1993

[4] C.H. The and R.T. Chin, " On the detection of dominant points on digital curves," IEEE Tr. On Pattern Analysis and Machine Intelligence, Vol.11, No.8, pp.859-872, 1989

[5] Y. Tseng, et al., " Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm" Pattern Recognition 20, pp. 791-806, 1999

[6] 이성훈, 조규태, 김진식, 김진형, 정철곤, 김상균, 문영수, 김지연, "비분할 자소 인식과 분할 자소 인식 결합을 통한 비디오 한글 인식", 제18회 영상처리 및 이해에 관한 워크샵, pp. 148-153, 2006