다양한 서식 문서에서 선에 의해 훼손된 문자열 복원

이창현°, 이관용, 김경환†, 최영우‡, 이일병
연세대학교 컴퓨터과학과, †서강대학교 전자공학과, ‡숙명여자대학교 전산학과

Restoration of Character String Stained with Line in Various Kinds of Form Documents

Changhyun Lee, Kwanyong Lee, Gyeonghwan Kim†, Yeongwoo Choi‡, Yillbyung Lee
Dept. of Computer Science, Yonsei University
†Dept. of Electronic Engineering, Sogang University
‡Dept. of Computer Science, Sookmyung Women's University

요 약

현 사회에서 사용하고 있는 문서들은 양식을 가지고 있는 서식문서가 대부분이며, 이러한 양식을 가지고 있는 서식문서는 사회가 발전함에 따라 전자 문서로의 변경이 요구되고 있다. 그러나 서식문서를 전자 문서로 바꾸는 작업은 쉬운 일이 아니며, 이 작업을 위해 문자인식 기술이 요구된다. 특히 서식문서에서 문자의 인식률을 높이기 위해서는 문서양식의 라인과 겹쳐진 문자에 대하여 라인제거 및 문자복원이 필수적이며, 또한 대부분의 서식 문서의 양식에 기입하는 내용은 문자열로 구성되어 있으므로 문자복원에 있어서 낱자 단위의 문자복원이 아닌 문자열 단위의 문자복원이 필요하다. 본 논문에서는 다양한 서식문서에서 라인과 겹쳐진 문자 영상에 대해 문자열 단위의 라인제거 및 복원하는 방법을 제안한다.

1. 서론

우리는 생활 전반에 걸쳐 형식이 있는 문서를 접하고 있으며 특히 은행전표나 신용카드 매출 전표와 같은 형식문서들은 처리의 중요성과 처리양의 증가로 관심이 집중되고 있다.

형식문서는 형식이 일정하다는 점에서 자동처리가 가능하나 고도로 발전된 현대사회에서도 형식문서의 완전 자동처리에는 많은 문제점들을 가지고 있다. 특히 문서 영상에서 추출해야 할 문자가 서식문서의 틀을 이루는 선에 의해 훼손된 경우에는 이를 효과적으로 복원할 수 있는 방법이 필요하다. 이러한 문제점들을 인식해서 많은 연구들이 수행되어졌으며 지금도 연구되고 있다.

선과 겹쳐진 상태로 인식을 수행할 경우에는 인식해야할 대상의 변형이 많아져 좋은 인식기로도 높은 인식률을 얻을 수 없으며 선과 문자가 접촉된 부분에서 단순히 자르는 방법은 추출되는 문자가 많이 훼손되기 때문에 역시 높은 인식률을 기대할 수 없다. 수학적 모폴로지 연산을 이용해서 라인을 제거하는 방법은 구현이 간단하나 라인제거로 인해 끊어진 문자 부분에 대한 복원이 부정확한 단점이 있다. 구조적인 방법을 이용하여 인식기와 연결한 복원방법은 처리시간의 증가로 비효율적이며 인쇄된 문서 양식을 색상 정보를 이용하여 드롭아웃 시키는 방법은 사용되는 데이터의 저장공간 문제로 잘 사용하고 있지 않다.

라인제거 및 복원순서의 관점에서 보았을 때 선을 먼저 제거한 후 훼손된 문자를 복원하는 방법은 선 제거시 문자의 일부도 제거되므로 문자 영상에 대한 많은 정보 손실로 문자 영상의 복원 시에 어려움이 있다. 선을 제거하기 전에 복원하는데 필요한 정보를 얻고, 이 정보를 이용하여 복원하려는 영역에서 문자부분을 추정하여 문자영역 이외의 영역을 제거하는 방법은 영상의 훼손을 최소화하면서 높은 인식률을 얻을 수 있다[1].

본 논문에서는 서식 문서의 양식 라인과 겹쳐진 문자열을 낱자 단위로 쪼개지 않고 문자열에 대한 기본적인 정보를 이용하여, 복원 영역에서 선이라고 추정되는 부분만을 제거하는 문자 영상복원 방법을 제안한다. 제안하는 복원 방법은 선과 문자가 접촉한 모든 유형을 따로 분류하지 않고도 숫자, 영문자, 한글 등의 인쇄체 및 필기체에 좋은 복원 결과를 보였다.

2. 제안하는 영상 복원 방법

2.1 라인 위치 검출

대부분의 선 검출 방법은 히스토그램에 적당한 임계값을 이용하여 수평선분과 수직선분을 검출하였다. 그러나 이러한 선 검출 방법은 점선이나 약간 기울어진 선분을 검출하는데 많은 문제점을 가지고 있으며 이를 보완하기 위해서는 많은 계산량을 필요로 하는 허프 변형을 이용하여 왔다.

본 논문에서는 선 위치를 검출하는데 많은 시간을 들여야 한다는 단점을 보완하기 위해 서식 처리 단계에서 넘어오는 기본적인 정보인 라인의 개수 정보를 히스토그램 정보에 사용하여 라인의 검출에 불필

본 논문은 정보통신부 '산·학·연 공동기술개발사업'의 지원을 받음.

요한 처리 시간을 줄였다.

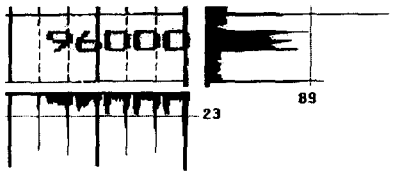


그림 1 개수 정보를 이용한 라인 위치 추정

2.2 라인의 두께 추정

라인을 정확히 제거하기 위해서는 라인의 두께 정보를 구하는 것이 필수이며 수직선과 수평선의 두께가 각각 다를 경우 수직선과 수평선의 두께를 따로 검출하여 정확한 라인의 굵기에 대한 정보를 얻었다. 형식 문서에서 많이 나타나는 같은 라인 안에서의 가변 두께를 가지는 라인의 두께 측정은 가변 두께가 많이 나타나는 수평선에 대하여 3개 영역으로 분할한 후 각 영역에 대하여 가장 빈도수가 높은 굵기를 각 영역에 대한 라인의 굵기로 사용하는 영역 분할에 의한 두께추정 방법을 사용하였다

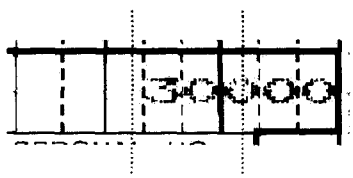


그림 2 영역분할에 의한 width 추정

2.3 문자의 획 두께 추정

라인으로 추정되는 부분을 제외한 영상에서 가로 세로 런-길이를 구한 후 이 값 중 빈도수가 가장 높은 값을 문자의 획 두께로 추정하였다.

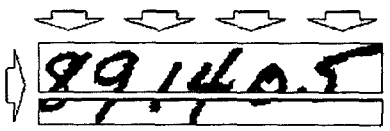


그림 3 라인을 제거한 문자부분의 가로 세로 런-길이 측정

2.4 접촉 검출

라인이라고 추정되는 위치에서의 수직 런-길이를 구하여 그 값이 라인의 두께보다 큰 경우 접촉되었다고 추정하고 이 정보를 후보로 등록한다. 이렇게 등록된 후보를 선별하기 위해서 등록된 위치에서의 수평 런-길이를 구하여 문자의 획 두께보다 3배정도 크거나 1, 2 픽셀 정도 크기를 갖고 수직 런-길이가 라인의 두께보다 1, 2 픽셀정도를 경우 등록된 후보는 노이즈로 간주하여 접촉되었다고 추정한 후보에서 제거한다.

2.5 라인 구성요소 세분화

라인 구성요소를 세분화하는 것은 복원을 위한 전 단계로서 라인을 구성하고 있는 요소들로 세분화하여 순수 라인부분은 제거하고 문자와 라인이 접친 라인 부분은 기본적인 몇 가지 유형으로 분리하여 각각 복원을 수행한다.

문자 접촉 검출단계에서 측정한 라인이라고 추정되는 위치에서의 수직 런-길이를 이용하여 접촉 유형을 4가지로 분류한다. 4가지 유형은 각각 위로 긴 경우, 아래로 긴 경우, 위 아래로 긴 경우, 수직 런-길이가 라인의 두께와 같은 경우이다.

이 4가지 유형중 위로 긴 경우와 아래로 긴 경우에 대해서 복원방법을 고려하여 런-길이의 크기를 기준으로 접촉 유형을 분리하면 수직 런-길이가 라인 두께보다 크고 라인 두께와 문자 획 두께를 더한 값보다 작거나 같은 접촉 유형과 수직 런-길이가 라인 두께와 문자 획 두께를 더한 값 보다 큰 경우 이렇게 2가지로 분류된다.

수직 런-길이가 라인 두께보다 크고 라인 두께와 문자 획 두께를 더한 값보다 작거나 같은 접촉 유형은 문자와 라인이 접한 형태이거나 문자가 라인에 일부 포함된 경우이다.

수직 런-길이가 라인 두께보다 작거나 같은 경우는 문자가 라인에 완전히 포함된 경우 혹은 문자가 라인과 접촉하지 않은 경우다.

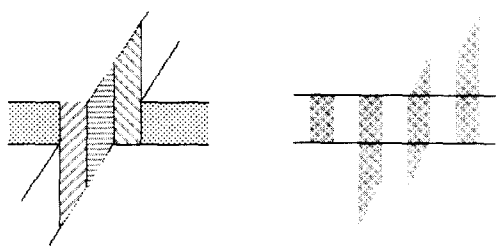


그림 4 수직 런-길이를 이용한 접촉유형 4가지

2.6 문자 복원

복원은 문자 영상의 일부분과 선의 일부분이 함께 존재하는 영역을 찾아 이들을 분리해서 문자 영상의 일부분만 추출하고 선의 일부분은 제거하는 작업이다.

위 아래로 긴 경우의 접촉 형태는 복원하려는 부분이 문자의 일부분만 존재하고 선의 일부분이 존재하지 않는 영역이기 때문에 복원작업을 할 필요가 없다.

위로 긴 경우와 아래로 긴 경우의 수직 런-길이가 라인 두께보다 크고 라인 두께와 문자 획 두께를 더한 값보다 작거나 같은 유형의 복원 방법은 문자 획 두께 정보를 이용하여 수직 런-길이에서 문자 획 두께 길이 만큼을 뺀 나머지 라인 부분을 제거하는 방법으로 복원한다.

위로 긴 경우와 아래로 긴 경우의 수직 런-길이가 라인 두께와 문자 획 두께를 더한 값 보다 큰 유형의 복원 방법은 라인의 두 변 부분과 문자의 획이 만나는 두 점을 잇는 직선의 방정식을 구하여 그 직선을 중심으로 문자 영상의 일부분과 그렇지 않은 부분을 분리하는 방법으로 복원한다.

수직 런-길이가 라인 두께보다 작아 문자의 획이 라인에 완전히

포함되거나 문자의 획이 라인과 접촉하지 않은 경우는 모호성 검출단계에서 처리한다.

2.7 모호성 검출

문자의 획이 라인에 완전히 포함된 경우와 문자의 획이 라인과 접촉하지 않은 경우에 수직 런-길이가 모두 라인 두께보다 작게 나타나기 때문에 수직 런-길이 만으로는 위의 2가지 유형을 구별할 수 없다. 이 문제를 해결하기 위해 문자의 구조적인 정보, 문자 획의 연속성, 주변의 접촉유형 등을 이용하여 문자가 라인에 완전히 포함되었다고 추정되는 곳에는 모호한 영역을 복원한 영상과 복원하지 않은 영상을 후보로 등록하여 신뢰도를 높이는 방법을 사용하였다.

3. 실험 환경 및 데이터

3.1 실험 환경

제안하는 선에 의해 훼손된 문자열 복원 시스템의 구현 언어로는 Visual C++ 6.0을 사용하였으며 Pentium 300 MHz 메모리 64MB의 PC에서 구현하고 실험하였다.

3.2 평가 대상 영상과 평가 기준

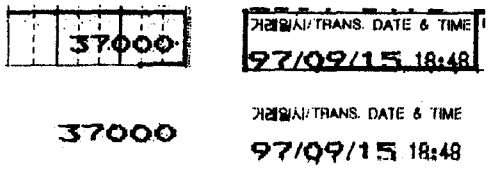
기울기 보정된 200dpi 이진 영상을 사용하였으며, 라인의 정보인 라인의 개수가 전 단계의 모듈로부터 주어진다고 가정하였다.

필기 한글, 숫자 열을 대상으로 하는 데이터는 시중은행 5곳(국민은행, 보람은행, 신한은행, 외환은행, 하나은행)의 입출금 전표를 사용하였으며, 인쇄된 숫자 열을 대상으로 하는 데이터는 신용카드의 매출전표(이지체크, 한국부가통신, 키스체크, 유공)로 하였다. 그리고 보다 객관적인 방법으로 제안하는 방법의 성능을 평가하기 위해서 NIST DB special 3번에서 추출한 2~10개의 낱자를 포함하는 500개의 숫자열 데이터에 대하여 실험하였다. 500개의 숫자열은 2312개의 숫자를 포함한다.

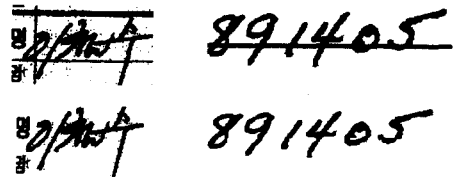
NIST DB의 원본 숫자열과 원본 숫자열에 무작위로 생성한 선을 위치 시킨후 제안하는 방법으로 복원한 숫자열의 인식 실험 결과를 비교하여 제안한 방법의 타당성을 검토했다.

4. 실험 결과

4.1 인쇄체 문자 영상



4.2 필기체 문자 영상



4.3 NIST DB에 대한 실험 결과

NIST DB의 원본 숫자열과 원본 숫자열에 무작위로 생성한 선을 위치 시킨후 제안하는 방법으로 복원한 숫자열을 인식 실험한 결과는 500개의 숫자열, 총 2312개의 숫자에 대하여 원본 숫자열은 85.21%의 인식률을 얻었고 접촉 복원된 숫자열은 83.35%의 인식률을 얻었다. 선에 접촉된 숫자열을 복원하지 않고 인식할 경우는 숫자열에 대한 분할이 제대로 이루어지지 않아 낮은 인식률을 보였다.

5. 결론 및 향후 연구 과제

본 논문에서는 서식 문서의 양식 라인과 겹쳐진 문자열을 낱자 단위로 쪼개지 않고 문자열에 대한 기본적인 정보를 이용하여, 복원 영역에서 선이라고 추정되는 부분만을 제거하는 문자 영상복원 방법을 제안하였다.

본 논문에서 제안하는 방법은 문자의 획과 선의 접촉 상태 그리고 라인의 개수와 같은 기본적인 정보만으로도 접촉된 문자열이 숫자, 영문자, 또는 한글 중 어느 문자집합에 속하든 관계없이 문자열을 원래 영상과 매우 가깝게 복원했다. 또한 선과 문자가 접촉한 모든 유형을 따로 분류하거나 정의하지 않아도 된다는 이점이 있다. 앞으로 라인과 문자열의 획이 완전히 포함되어 복원하기 어려웠던 부분의 영상 훼손을 최소화하여 복원의 신뢰도를 높이는 방법을 개선하고 보완한다면 더욱 좋은 결과를 얻을 수 있을 것이다.

참고문헌

[1] Youngtae Chung, Kwanyong Lee, Jonghyun Paik, Yillbyung Lee, "Extraction and Restoration of Digits Touching or Overlapping Lines," Proceedings of the 13th IAPR International Conference on Pattern Recognition, Vol 3, pp.155-159, 1996.

[2] 심상옥, 권영빈, "형식문서상의 필드추출 및 접촉문자 복원", 1998년도 한국정보과학회 봄 학술발표논문집, 25권 1호, pp.701-703, 1998.

[3] Y.H.Tseng, H.J.Lee "Interfered-character Recognition by Removing Interfering-lines and Adjusting Feature Weights", Proceedings 14th International Conference on Pattern Recognition, Vol 2, pp.1865-1867, 1998