Space Coding 을 이용한 Text Watermarking

황 미 란, 추 현 곤, *최 종 욱, 김 회 율
한양대학교 전자통신전파공학과, *(주)마크애니
전화: 02-2290-0561 / 핸드폰 : 011-9795-9860

Text Watermarking using Space Coding

Mi-Ran Hwang, Hyon-Gon Choo, Jong-Wook Choi, Whoi-Yul Kim
Division of Electrical and Computer Engineering, Hanyang University,
MarkAny Research Institute
E-mail : mirang@vision.hanyang.ac.kr

Abstract

In this paper, we propose a new text watermarking method using space coding and PN sequence. A PN sequence generated from user message modifies the space between words in each line. The detection can be done without original text image using the average space with in the text. Experimental results show that proposed method has the invisible property and robustness to the attack such as the elimination of words in the text.

I. 서론

인터넷과 통신망의 보급으로 인한 컴퓨터의 일반화와 스캐너와 레이저 프린터의 인쇄 기술의 향상은 텍스트 문서에 대한 위. 변조 및 신빙성에 대한 보완 기술을 요구하게 만들고 있다. 텍스트 워터마킹은 문서 내에 사람이 느끼지 못하는 새로운 정보를 추가하여, 위-변조 방지, 전자문서(e-Book)의 저작권 표시 등 여러 가지 응용에 사용될 수 있는 기술이다. 기존의 텍스트 워터마킹 방법은 크게 다음과 같다[1-3].

1) *Line-Shift Coding*: 텍스트 문서의 줄 위치를 수직 방향으로 이동하여 정보 삽입
2) *Word-Shift Coding*: 텍스트 문서의 줄을 구성하는 단어의 위치를 수평 방향으로 이동하여 정보 삽입
3) *Feature Coding*: 텍스트 문서의 주요 특징 요소를 선택하고 변화시켜 정보 삽입

Line-shift coding 과 word-shift coding 은 워터마크 정보를 찾는 과정에서 워터마크 정보가 삽입되지 않은 원본 영상이 필요하며, 영상에서의 변화량과 워터마크를 구별하는 기준이 분명히 있어야 하는 단점이 있다. Feature coding 방법은 특정 feature 에 삽입된 잡음에 의해 쉽게 영향을 받기 때문에 많은 제약이 따른다. Huang 은 이런 단점을 보완하기 위해 inter-word space 를 이용한 텍스트 워터마킹 방법을 제안하였다. Huang 의 방법은 문서내의 단어 사이의 여백을 sine curve 의 크기 성분에 따라 변화시키는 방법을 사용하였다. 그러나, 이 방법은 sine curve 를 사용함으로 인하여, 에러나 공격에 강인하지 못한 속성을 가지게 된다[4].

본 논문에서는 space coding 을 사용한 새로운 텍스트 워터마킹 방법을 제안한다. 텍스트 문서에서 워터마킹 정보를 삽입하고자 하는 라인의 개수에 따라 적당한 PN(Pseudo Noise) 코드를 발생시켜, 한 라인에서의 단어 사이의 여백 간격이 PN 코드 성질을 갖도록 단어를 재배열하게 된다. 제안된 방법은 워터마크 삽입 영역이 일정한 구간의 평균 공간 값을 사용하므로 눈에 띄지 않는 특성을 갖고 외부 공격에 의한 왜곡에 강인한 특성을 갖는 워터마킹 방법으로 사용될 수 있다[5].

본 논문의 구성은 다음과 같다. II 장에서는 space coding 과 PN 코드를 이용한 워터마킹 삽입 과정을 살펴보고, III 장에서는 삽입된 워터마킹 정보의 검출 과정을 설명한다. IV 장에는 제안된 방법에 대한 실험결과를 분석하고 V 장에서 결론을 내린다.

II. Watermark Embedding

제안된 방법은 워터마크 정보를 삽입하는데 있어서 space coding 을 기본으로 하는 PN 코드를 사용한다. Space coding 이란 텍스트 문서에 존재하는 단어 내의 문자 사이 여백을 변화시켜 단어 사이 여백 간격을 조절하는 방법이다.

그림 1 에 나타난 워터마크 정보를 입력하는 과정은 다음과 같다. 영상에 대한 수직, 수평 projection profile 정보를 이용하여 각 라인 및 단어를 분리한 후, 라인 별로 단어 사이의 평균 여백을 계산한다. 이 평균 여백에 대한 정보를 이용하여 단어의 문자간 여백 간격을 변화시켜 워터마크 정보를 삽입한다.

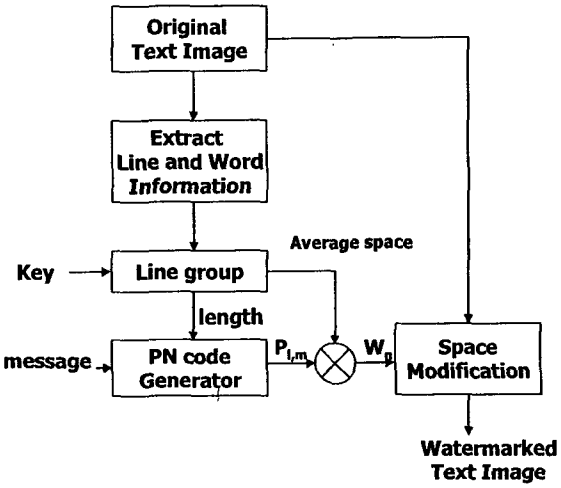

그림 1. Watermarking Embedding Procedure

2.1 Space information extraction

이진 데이터를 갖는 텍스트 문서는 글자에 해당하는 픽셀과 배경에 해당하는 픽셀로 이루어져 있으므로 글자에 해당하는 픽셀을 수직, 수평 projection profile 을 이용하여 텍스트 단어 사이의 여백 정보를 얻을 수 있다.

먼저 수평 방향 projection 을 통해 얻은 profile 정보로 텍스트를 이루는 각 라인의 위치 정보를 알아낸다. 수평 projection 정보로부터 텍스트 라인의 t(top), b(bottom)을 구할 수 있고, 이 정보를 바탕으로 라인 별로 수직 projection 하여 각 라인의 단어의 개수와 위치 정보를 알 수 있다.

$$v(x) = \sum_{y=t}^{b} f(x,y)$$
$$where \quad f(x,y): text\ image \qquad (1)$$
$$v(x): vertical\ projection$$

이와 같이 얻어진 단어의 위치 정보를 통해 단어 사이의 여백 정보를 계산할 수 있다. 그림 2 는 이 수평, 수직 projection profile 을 이용하여 단어를 구분한 예이다.

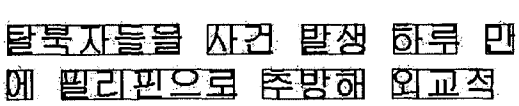

그림 2. 단어 구분의 예

2.2 Space coding 을 이용한 워터마크 삽입

워터마크 정보를 삽입하기 위해 입력 받은 키 정보로부터 삽입할 라인을 선택하게 된다. 키 값보다 많은 단어 개수를 갖는 라인을 선택하여 하나의 그룹으로 묶고, 그룹 내의 라인의 개수와 사용자 메시지를 이용하여 하나의 PN 코드 시컨스를 생성한다 이 시컨스의 각각의 코드 값은 라인 그룹의 특징에 비례하여 변형된다.

라인의 특징으로 사용되는 정보는 삽입할 라인 내에 여백의 평균 길이 $\overline{S}$ 이며 다음과 같이 구할 수 있다.

$$\overline{S} = \frac{\overline{S}_t}{(d-1)}, \quad d \neq 1 \qquad (2)$$

여기서 $\overline{S}_t$ 는 임의의 라인의 단어 간 여백의 총 합이며, d 는 라인의 단어 개수이다. 그림 3 은 임의의 라인에 대한 단어간 여백 평균과 단어 개수 분포를 보여준다.

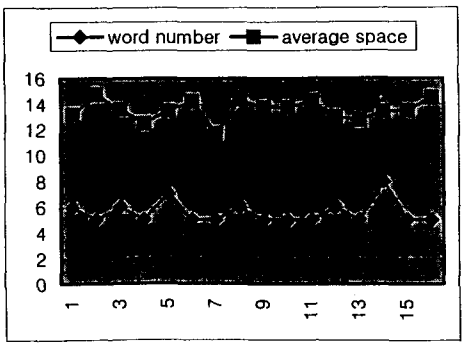

그림 3. 각 라인의 단어 개수와 여백 평균

텍스트 영상에 워터마크를 삽입하는 방법은 다음과 같다.

1) 텍스트 내에서 입력 받은 키 정보와 비교하여 단어의 개수가 키보다 많은 라인을 하나의 그룹으로 묶는다.
2) 그룹 내의 라인에 대해 각각 평균 여백 $\overline{S}$ 을 구한다.
3) 각 라인의 $\overline{S}$ 로부터 그룹 전체의 단어간 여백 평균을 구한다. N 개의 라인으로 이루어진 그룹 전체에 대한 단어간 여백 평균 a 는 다음과 같다.

$$a = \frac{\sum^{N} \overline{S}_n}{N} \tag{3}$$

4) 입력 받은 메시지로부터 PN 코드 시컨스를 생성한다. 시컨스의 길이는 그룹 내의 라인 개수에 따라 정해진다. 즉, 그룹 내의 라인 개수에 따른 시컨스의 길이가 l 이며 메시지 m 일 때에 대한 Pseudo Noise 코드 시컨스 $P_{l,m}$ 을 발생시킨다. 발생시킨 코드의 각 비트별 값은 적절한 세기 C 와 그룹 여백 평균 a 만큼의 크기 값으로 변화된다.

$$W_n = C \times a \times P_{l,m}(n)$$
$$where\ n : bit\ number \tag{4}$$

5) 그룹에 대한 여백 평균 a 와 삽입할 워터마크 정보 W_n 를 더하여 각 라인의 새로운 단어간 여백 평균을 계산한다.

$$\overline{S}_n' = a + W_n \tag{5}$$

6) 5)의 계산을 통해 얻어진 여백 평균값을 반영하기 위해, 각 라인에서의 단어 사이의 여백 간격을 수정한다. 실제로 단어간 여백 간격을 수정하기 위해, 여백을 직접 수정하지 않고, 단어 자체의 길이를 변화시켜 워터마크 정보를 텍스트 이미지에 분산 시킨다. 각 단어 별 변화량 Δ_i 는 다음과 같이 구할 수 있다.

$$\Delta_i = \left| \overline{S}_c \times \frac{L_i}{\sum_{i=1}^{d} L_i} \right|, \quad \overline{S}_c = \overline{S}_n' - \overline{S}_n \tag{6}$$

여기서 $\overline{S}_c$ 는 각 라인 별 워터마크 정보가 삽입된 후와 원본에서의 여백 평균 차이이며, $\overline{S}_c$ 를 라인 내의 각 단어 길이에 비례하도록 나눈 값이 Δ_i 이다. 이때 단어의 픽셀 길이를 변형하기 위해서 단어를 이루는 문자 사이의 여백 간격을 조절한다.

III. Watermark Detecting

워터마크를 추출하는 과정은 삽입하는 과정과 유사하다. 삽입할 때 사용된 키를 이용하여 워터마크 정보가 삽입된 라인들을 선택하고 이 라인들을 그룹으로 묶어준다. 워터마크 정보가 삽입된 이미지로부터 선택한 라인 그룹에 대하여 n 번째 라인의 단어간 여백을 $Y(n)$ 이라고 하면 라인 그룹의 단어간 여백 평균 a 와의 차로 워터마크 정보에 해당하는 PN 코드 시컨스를 추출할 수 있다.

$$W_n = Y(n) - a \tag{7}$$

식 (9)와 같이, 추출된 코드 정보 W_n 과 입력 받은 키로 생성되는 참조(reference) 코드 $P_{i,j}$ 정보와의 유사도(correlation) 비교를 통해 워터마크 정보를 얻을 수 있다.

$$r(j) = \frac{1}{N} \sum^{N} W(n) P_{l,j}$$
$$where\ r(j) : correlation\ value \tag{8}$$
$$l : group\ length$$
$$j : random\ seed$$

IV. 실험 및 결과

4.1 실험 환경

제안한 방법의 성능 효율성을 나타내기 위해 Huang 의 방법과 비교하였다. 실험을 통해 임의로 문서 내에 단어를 변경, 이동하거나 삭제하는 경우에도 워터마크 정보가 얼마나 정확히 추출 되는가를 살펴본다. 실험에 사용된 텍스트 이미지는 다양한 폰트와 크기를 갖는 이진 텍스트 이미지를 사용하였다.

4.2 실험 결과

그림 4 는 제안된 방법을 이용하여 워터마크를 삽입한 결과이다. 그림 4 의 결과를 통해 워터마크가 입력된 영상이 원본 영상과 비해 인지적으로 차이가 없음을 알 수 있다.

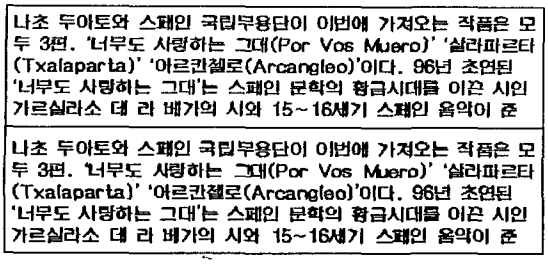

그림 4. 제안된 방법을 이용하여 워터마킹 된 영상: (원본 영상: 위, 워터마킹 된 영상: 아래)

표 1 은 워터마크를 삽입한 영상에 대한 검출률 실험 결과이다. Sine curve 형태의 워터마크 정보보다 PN 코드 형태의 워터마크 정보를 작은 강도로 삽입하였을 때, 정보의 손실에 더 강인함을 확인할 수 있다.

표 1. 삽입된 워터마크에 대한 검출률 비교

	1	2
Original	85% (17/20)	95% (19/20)
1 Word Erase	85% (17/20)	100% (20/20)
3 Word Erase	80% (16/20)	100% (20/20)

(1: Huang 의 방법, 2: 제안된 방법)

V. 결론

본 논문에서는 텍스트의 단어와 단어 사이의 여백에 대한 space coding 을 이용하여 PN 코드 형태의 워터마크 정보를 삽입하는 방법을 제안하였다.

사용자의 메시지에 따라 생성된 PN 코드는 텍스트 문서 내의 라인에 있는 단어 사이 여백 간격을 변화시켜 삽입된다. 워터마크 정보의 추출은 텍스트 문서 내의 라인에 대한 평균 여백 정보를 통해 가능하다. 제안된 방법은 워터마크의 정보를 추출하는 데 원본 영상을 요구하지 않으며 실험을 통하여 제안된 방법이 사람 눈에 덜 민감하게 정보를 삽입할 수 있으면서, 단어를 삭제하거나 변경하는 경우에도 정확하게 워터마크 정보를 추출함을 확인할 수 있다.

추후 연구 방향으로는 라인 단위로 삽입이 가능한 워터마크의 정보량을 증가시키는 방법에 대한 연구가 필요하다.

참고문헌

[1] S. H. Low, N. F. Maxemchuk, "Performance Comparison of Two Text Marking Methods," *IEEE J. Select Areas Commun.*, vol. 16, May 1998.

[2] S. H. Low, N. F. Maxemchuk, J. T. Brassil, and L. O'Gorman, "Document marking and identification using both line and word shifting," in *Proc. INFOCOM'95*, Boston, MA, Apr. 1995.

[3] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, "Document identification for copyright protection using cancroid detection," *IEEE Trans. Communications*, vol. 46, Mar. 1998.

[4] D. Huang and H. Yan, "Interword distance changes represented by sine waves for watermarking text images," *IEEE Trans. on Circuits and Systems for Video Technology*, in press.

[5] R. E. Ziener and W. H. Transter, *Principles of communications*, 2nd Ed. New York: JHON WILEY & SONS, Inc., 1995.