

영상에 기반한 음란 유포 사이트의 차단

조동욱* 최병갑**

*: 충북과학대학 정보통신학과 **:목원대학교 컴퓨터공학과

e-mail : ducho@ctech.ac.kr

Blocking of Harmful Internet Sex Sites Based on Image Processing

Dong Uk Cho* Byung Kap Choi**

* : Chungbuk Provincial Univ, of Science & Technology, ** : Mokwon University

요 약

음란 유포 사이트는 70만개에 육박하고 있으며 초등학생들에게도 매일 평균 2개씩의 음란 사이트 안내에 대한 메일이 가고 있는 등 이에 대한 사회적 대응이 필요한 시점이 되었다. 특히 한글로 된 음란 유포 사이트가 세계 2위이며 그 증가율은 세계 1위인 상황이기 때문에 더 더욱 이를 차단하기 위한 기술적 방법론이 개발되어야만 하는 실정이다. 본 논문에서는 이를 위해 신호처리기법에 기반한 음란 유포 사이트 차단 방법을 제안하고자 한다. 전체 시스템중 본 논문은 영상에 기반하여 음란 유포 사이트를 차단하는 방법을 제안하고자 하며 실험에 의해 본 논문의 유용성을 입증하고자 한다.

을 입증하고자 한다.

1. 서론

인터넷 역기능에 대한 심각성이 날로 심해지고 있다[1]. 지난 1월 워바이러스에 의한 인터넷 대란[2]으로부터 시작해서 자유게시판의 부적절한 글쓰기 형태, 채팅으로 인한 청소년들의 탈선과 주부들의 탈선 그리고 국적 불명의 전자 우편 사용 어휘등에 대한 대책이 시급한 실정이다. 하루의 일과를 스팸메일 지우기로 시작하는 현재의 생활 패턴에서 제일 문제가 심각한 것은 음란 유포 사이트가 스팸메일로서 학령인구의 청소년들에게 배포되는 것으로써 이는 바이러스 메일 유포와 더불어 반드시 기술적으로 해결해야 할 과제가 아닐수 없다. 이를 위해 본 논문에서는 신호 처리 기법에 기반한 음란 유포 사이트 차단 방법을 제안하고자 한다. 신호처리에 기반한 방법을 적용한 이유는 현재와 같이 단어 기반과 목록 기반에 기초한 방법으로는 차단이 되지 않는 음란 사이트가 존재하고 있으며 특히 유료사이트가 아니라도 이와 비슷한 유형의 무료 사이트 자료들이 인터넷상에 상당히 있기 때문이다. 이를 위해 신호처리 기반에 기반한 방법을 새로이 제안하고자 하며 특히 본 논문은 전체 신호 처리 기법에 기반한 시스템 중 영상에 기반한

2. 영상에 기초한 유포 음란 사이트 차단

2.1 영상 분할

영상분할을 위해 텍스처 분석을 행한다. 텍스처 분석을 행하는 방법은 크게 구조적 방법[3],[4], 신호처리 방법 [5],[6]과 통계적 방법[7]~[8]으로 구분된다. 본 논문에서는 텍스처 분석을 통계적 방법인 GLCM 방법중 변화에 민감하게 반응하는 모멘트와 동일 성질에 반응하는 동차성을 조합하는 방법[20]을 참조하여 빈도수를 통한 GLCM내의 값들의 정규화와 동차성의 정규화를 행하는 방법을 제안하여 신체의 주요 성적 부분의 영역을 추출하고자 한다. 우선 GLCM(Gray Level Co-occurrence Matrix) $G[i, j]$ 는 변위벡터 $d = (dx, dy)$ 를 결정하고 명도 i 와 j 를 갖고 d 만큼 떨어져 있는 화소들의 쌍을 계산하여 생성한다. 이때 전체 빈도수를 GLCM의 각 원소값들에 나누어줌으로써 $\sum G[i, j] = 1$ 이 되도록 조정한다. 따라서 이를 이용하여 모멘트 M 은 하식 (1)과 같이 구한다.

방법을 제안하고자 하며 실험에 의해 본 논문의 유용성

$$M = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-j)^2 G[i, j] \quad (1)$$

$$\sum_{k=0}^n a_k \sum_{i=0}^M x_i^{j+k} = \sum_{i=0}^M y_i x_i^j \quad (6)$$

또한 정규화된 동차성 H 는 하식 (2)와 같이 새로이 정의하여 그 값을 구한다.

$$H = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{255 - |i-j|}{255} G[i, j] \quad (2)$$

이를 4×4 윈도우 크기로 연속 수행하여 두 개의 배열 $[A]$ 와 $[B]$ 에 각각 저장한다. 이때 배열 $[A]$ 와 배열 $[B]$ 에 대해 특정한 임계치의 값으로 임계치 작업을 행하며 이를 통해 배열 $[A]$ 와 배열 $[B]$ 에 대해 OR 작업을 수행한다.

이의 결과는 신체의 주요 성적 부위에 해당하는 영역이 되며 이를 Y축으로 프로젝션을 행하여 차후 형성된 히스토그램에 대한 분석을 행한다.

2.2 곡선적합

여성과 남성의 신체의 주요 성적 부위는 2차곡선이나 일량분포로 나타낼 수 있으므로 이에 대한 곡선적합을 수행한다.

$(M+1)$ 개의 자료 $\{(x_i, y_i) | i = 0, 1, \dots, M\}$ 를

$n (< M)$ 차의 다항식 $P_n(x) = \sum_{k=0}^n a_k x^k$ 에 근사시키는 곡선적합은 아래와 같다.

$$E = \sum_{i=0}^M (y_i - P(x_i))^2 \quad (3)$$

위 식을 최소화하는 a_0, a_1, \dots, a_n 을 구하기 위해서는 (3) 식에서

$$\begin{aligned} E &= \sum_{i=0}^M y_i^2 - 2 \sum_{i=0}^M P(x_i) y_i + \sum_{i=0}^M (P(x_i))^2 \\ &= \sum_{i=0}^M y_i^2 - 2 \sum_{i=0}^M \left(\sum_{k=0}^n a_k x_i^k \right) y_i + \sum_{i=0}^M \left(\sum_{k=0}^n a_k x_i^k \right)^2 \\ &= \sum_{i=0}^M y_i^2 - 2 \sum_{i=0}^M a_i \left(\sum_{k=0}^n y_i x_i^k \right) + \sum_{i=0}^M \sum_{k=0}^n \sum_{l=0}^n a_i a_k \left(\sum_{j=0}^n x_i^{j+k} \right) \end{aligned} \quad (4)$$

(4)식을 최소화하려면 $j = 0, 1, \dots, n$ 에 대하여

$$\frac{\partial E}{\partial a_j} = \left(-2 \sum_{i=0}^M y_i x_i^j \right) + 2 \sum_{k=0}^n a_k \sum_{i=0}^M x_i^{j+k} = 0 \quad (5)$$

따라서 $(n+1)$ 개의 미지수 a_j 에 대하여

이차곡선에 대해서는

$$a_0 \sum_{i=0}^M x_i^0 + a_1 \sum_{i=0}^M x_i^1 + a_2 \sum_{i=0}^M x_i^2 = \sum_{i=0}^M y_i x_i^0 \quad (7)$$

$$a_0 \sum_{i=0}^M x_i^1 + a_1 \sum_{i=0}^M x_i^2 + a_2 \sum_{i=0}^M x_i^3 = \sum_{i=0}^M y_i x_i^1 \quad (8)$$

$$a_0 \sum_{i=0}^M x_i^2 + a_1 \sum_{i=0}^M x_i^3 + a_2 \sum_{i=0}^M x_i^4 = \sum_{i=0}^M y_i x_i^2 \quad (9)$$

따라서 a_0 는 아래와 같이 구할 수 있다.

$$a_0 = \frac{\begin{vmatrix} \sum y_i x_i^0 & \sum x_i^1 & \sum x_i^2 \\ \sum y_i x_i^1 & \sum x_i^2 & \sum x_i^3 \\ \sum y_i x_i^2 & \sum x_i^3 & \sum x_i^4 \end{vmatrix}}{\begin{vmatrix} \sum x_i^0 & \sum x_i^1 & \sum x_i^2 \\ \sum x_i^1 & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{vmatrix}} = \frac{|D_0|}{|D|} \quad (10)$$

마찬가지로 a_1, a_2 도 식 (11), (12)과 같이 하여 구할 수 있게 된다.

$$a_1 = \frac{|D_1|}{|D|} \quad (11)$$

$$a_2 = \frac{|D_2|}{|D|} \quad (12)$$

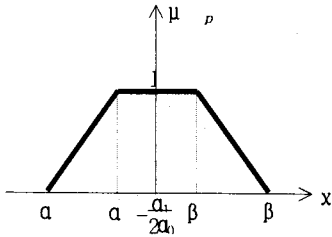
또한 일량분포에 대한 곡선적합은 $y = a + bx$ 에 대해 아래식과 같이 정리되어 일량분포에 대한 곡선적합을 수행할 수 있게 된다.

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{M \sum x_k^2 - (\sum x_k)^2} \begin{bmatrix} \sum x_k^2 & -\sum x_k \\ -\sum x_k & M \end{bmatrix} \begin{bmatrix} \sum y_k \\ \sum x_k y_k \end{bmatrix} \quad (13)$$

최종적으로 곡선적합 수식의 적용을 위해 아래와 같은 단계를 거친다. 우선 이차곡선 $f(x) = a_0 x^2 + a_1 x + a_2$ 에 대해 다음과 같은 작업을 수행한다. 이차곡선 $f(x)$ 의 이차항의 계수 a_0 가 $a_0 > 0$ 이면 곡선적합의 정합과정을 거치고 아닌 경우 의미 없는 영역으로 분류한다. 이때

$\frac{df(x)}{dx} = 0$ 이 되는 x 의 값 $-\frac{a_1}{2a_0}$ 가 중앙에 있는지 여부와 기울기의 완만한 정도가 주요 평가 척도가 된다.

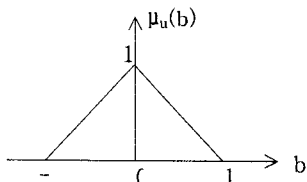
아래에 $x = -\frac{a_1}{2a_0}$ 이 중앙여부를 결정하는 함수식을 나타내었다.



(그림 1) 중앙여부를 결정하는 함수식

$$\begin{aligned} \mu_p(x) &= 1, & \frac{4}{5} * (\frac{\alpha+\beta}{2}) \leq x \leq \frac{6}{5} * (\frac{\alpha+\beta}{2}) \\ &= \frac{x-\alpha}{\alpha'-\alpha}, & \alpha \leq x \leq \alpha' \\ &= \frac{\beta-x}{\beta-\beta'}, & \beta \leq x \leq \beta' \end{aligned} \tag{14}$$

여기서 α 와 β 는 이차곡선의 시작과 끝점을 의미한다. 곡선적합이 일차식인 경우 $g(x) = a + bx$ 에서 기울기 b 에 대해 아래와 같은 함수식을 적용하여 일량분포의 정도(degree)를 나타내도록 한다.



(그림 2) 일량분포의 정도에 대한 함수식

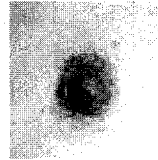
이때 좌표값 b 는 기울기이며 $\tan^{-1}(\theta)$ 값이 된다.

$$\begin{aligned} \mu_u(b) &= x+1 & -1 \leq b \leq 0 \\ &= 1-x & 0 \leq b \leq 1 \end{aligned} \tag{15}$$

3. 실험 및 고찰

본 논문에서의 실험은 IBM-PC 상에서 수행하였다. 우선 (그림 3)이 여성의 성적 부위 그리고 (그림 4)가 영역 분할 결과, (그림 5)가 히스토그램 분석 결과이다. 마찬가지로 (그림 6), (그림 9)가 여성의 성적 부위 그리고 이에 대한 영역 분할 결과가 (그림 7)과 (그림 10)이며 히스토그램 분석 결과가 각각 (그림 8)과 (그림 11)이다. 남성의 성적 부위에 대한 실험 결과는 (그림 12)~(그림 14)에 나타내었다.

아울러 <표 1>에 각각의 신체 부위에 대한 곡선적합을 수행한 결과를 나타내었다. 실험 결과에서 알 수 있듯이 본 논문에서 제안한 방법이 남, 여 주요 성적 부위에 대한 영역 분할과 히스토그램 분석 그리고 이에 따른 곡선적합을 올바르게 행할 수 있음을 확인할 수 있다. 차후 음향기반에 기초한 방법과 음변 논문에서 다룬 방법을 결합하여 목록기반과 단어기반에 의해 필터링 되지 않는 음란 유희사이트를 완전 차단하는 소프트웨어의 상용화에 착수해야할 것으로 여겨진다.



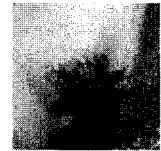
(그림 3) 여성의 성적 신체 부위



(그림 4) 영역 분할 결과



(그림 5) 히스토그램 (투영)



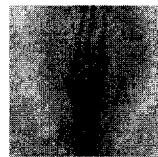
(그림 6) 여성의 성적 신체 부위



(그림 7) 영역 분할 결과



(그림 8) 히스토그램(투영)



(그림 9) 여성의 성적 신체 부위



(그림 10) 영역 분할 결과



(그림 11) 히스토그램(투영)



(그림 12) 남성의 성적 신체 부위



(그림 13) 영역 분할 결과



(그림 14) 히스토그램(투영)

<표 1> 곡선적합 결과

대상 영상	곡선 적합식
그림 5 (여성)	$f(x) = -10.9975x^2 + 1.7047x - 0.0260$
그림 8 (여성)	$f(x) = -9.4905x^2 + 2.0014x - 0.0244$
그림 11 (여성)	$f(x) = 15.4450 - 0.1126x$
그림 14 (남성)	$f(x) = 10.3231 - 0.1361x$

4. 결론

본 논문에서는 인터넷 역기능에 있어 가장 큰 사회적 문제로 대두되고 있는 음란사이트 차단 방법에 대해 제안하였다. 특히 청소년들을 위해 주니어 계정을 포털 사이트에서 따로 운영을 해도 이에 상관 없이 하루 2통의 음란사이트 안내 메일이 발송되는등 이에 대한 문제는 대단히 심각한 문제가 아닐수 없다. 이를 위해 기존의 방법들에서는 목록 기반 방법과 유해 단어기반에 기반하여 음란 사이트를 차단했는데 이 방법들은 새로운 내용을 목록기반 DB에 수시로 업데이트 해야 하는 문제점과 유해 단어가 주기적으로 교묘히 변경되는 문제로 말미암아 필터링이 올바르게 수행되지 않는 경우가 많았다. 본 논문에서는 이를 보완하기 위해 음향 방법과 영상기반에 기초하여 음란 유해사이트를 차단하는 신호처리 기반 방법을 제안하였으며 특히 본 논문은 전체 시스템 중 영상에 기반한 방법론에 대해 다루었다. 이를 위해 텍스처 분석에 의한 영상분할, 히스토그램 분석, 곡선적합 등을 적용하여 음란사이트를 차단하였으며 실험에 의해 제안한 방법의 유용성을 입증하였다. 향후에는 음향기반 방법과 결부하여 단어기반과 목록기반에서 필터링되지 않은 음란 유해사이트에 대해 완전한 차단이 이루어지는 상용 소프트웨어 개발이 이루어지도록 하기 위한 노력이 경주되어야 하리라 여겨진다.

참 고 문 헌

- [1] 디지털조선, 전세계 유해사이트 68만개, 2003년 4월 27일자
- [2] 조선일보, 인터넷대란, 2003년 1월 27일자
- [3] R. C Gonzalez and R. E. Woods, Digital Image Processing, Addison Wesley, 1992
- [4] N. Ahuja and A. Rosenfeld, "Mosaic Models for

Textures", IEEE Trans. on PAMI, 1981

[5] V. Alvestad, "Unsupervised Image Texture Segmentation with Optimized Filters", Master's thesis, stavanger, Norway, 1995

[6] F. Farrokhinic and A. K. Jain, "A Multichannel Filtering Approach to Texture Segmentation", Proc. of CVPR, 1991

[7] J. R. Parker, Algorithms for Image Processing and Computer Vision, John Wiley & Sons, pp. 150-175, 1997

[8] H. C. Lin et al, "Extracting Periodicity of a Regular Texture Based on Autocorrelation Functions", Pattern Recognition letters, 1997