이미지 블랙리스트 기반 저작권 침해 의심 사이트 탐지 기법

김의진*, 정인수**, 송유래**, 곽진***
*아주대학교 컴퓨터공학과 정보보호응용및보증연구실
,*아주대학교 사이버보안학과
dmlwls0403@ajou.ac.kr, jis0727@ajou.ac.kr, clara701@ajou.ac.kr, security@ajou.ac.kr

Detection Technique of Suspected Piracy Sites based on Image Black List

Eui-Jin Kim*, In-Su Jung**, Yu-Rae Song**, Jin Kwak***
*ISAA Lab., Dept. of Computer Engineering, Ajou University
,*Dept. of Cyber Security, Ajou University

요 약

저작권 콘텐츠의 해외 진출과 함께, 국내·외 저작권 시장 규모가 증가하고 있다. 이와 동시에 등장한 저작권 침해사이트는 메인 페이지에 저작권 침해사이트를 대표하는 이미지를 게시하는 특징이 있다. 이러한 저작권 침해사이트는 음악, 영화, 드라마 등의 저작권 콘텐츠를 불법 유통시키며 저작권 시장에 피해를 입히고 있다. 공공기관에서는 저작권 침해를 방지하기 위해 저작권 침해사이트를 차단하는 등의 대응을 하고 있지만, 저작권 침해사이트의 생성 속도에 비해 침해 여부 판단 속도가 상대적으로 느려서 차단에 어려움이 존재한다. 따라서, 본 논문에서는 저작권 침해사이트의 대표 이미지를 활용한 이미지 블랙리스트에 기반하여 저작권 침해 의심 사이트 탐지 기법을 제안하고자 한다.

1. 서론

국내 저작권 콘텐츠가 해외 시장으로 수출됨에 따라 국내·외 저작권 시장 규모가 증가하고 있다. 이에 따라, 저작물을 불법으로 유통하여 금전적 수익을 창출하는 저작권 침해사이트가 등장하였다[1,2]. 이러한 저작권 침해사이트가 음악, 영화, 드라마 등의 콘텐츠를 다양한 경로를 통해 불법 유통하고 있어, 저작권 시장은 직·간접적인 피해를 입고 있다[3]. 공공기관에서는 저작권 침해를 방지하기 위해 저작권 침해사이트를 차단하는 등의 대응을 하고 있지만, 침해사이트 생성 속도에 비해 침해 여부 판단 속도가 느려 실효성이 낮다[4]. 저작권 침해사이트는 해당 사이트를 대표하는 이미지를 웹사이트 메인 페이지에 게시하여 클라이언트에게 제공하는 특징이 존재한다[1,2]. 따라서, 본 논문에서는 저작권 침해사이트 침해 여부 판단을 위해 저작권 침해사이트의 대표 이미지를 활용한 이미지 블랙리스트를 기반으로 저작권 침해 의심 사이트 탐지 기법을 제안하고자 한다. 이때, 저작권 침해 의심 사이트는 저작권 침해 여부가 정확하게 판단되지 않고 침해가 의심되는 사이트를 말한다.

본 논문에서는 2장에서 관련 연구로 템플릿 매칭, 저작권 침해사이트의 특징과 링크 모음 사이트의 특징에 대하여 분석하고 3장에서는 이미지 블랙리스트 데이터셋에 대하여 설명한다. 4장에서 이미지 블랙리스트 기반 저작권 침해 의심 사이트 탐지 기법을 제안하고, 5장에서 결론을 맺는다.

2. 관련 연구

2.1 템플릿 매칭(Template Matching)

템플릿 매칭은 패턴 또는 템플릿과 가장 비슷한 부분을 이미지에서 정확히 판별하는 기술로, 패턴인식, 트래킹, 스테레오 매칭 등 컴퓨터 비전에서 사용되는 기법이다. 높은 빈도로 사용되는 템플릿 매칭 알고리즘으로는 Sum of Squared Differences (SSD) 와 Sum of Absolute Differences (SAD)가 있다. SSD와 SAD 방법은 간단한 계산 방법과 빠른 속도로 인해 많이 사용되지만, 노이즈에 취약한 단점이 있다[5].

본 연구는 문화체육관광부 및 한국저작권위원회의 2021년도 저작권연구개발사업의 연구결과로 수행되었음(2019-PF-9500).

2.2 저작권 침해사이트 특징

저작권 침해사이트는 저작물을 저작권자의 허가 없이 불법적으로 복사 및 전파하는 사이트를 말한다. 대표적인 저작권 침해사이트로는 토렌트, 영상 스트리밍, 웹툰 사이트가 있으며, 각 사이트 유형별로 특징이 존재한다.

토렌트 사이트는 저작물을 무료로 다운로드할 수 있는 마그넷, 다운로드 링크가 존재한다. 또한, 영상 스트리밍 사이트는 저작물을 이용할 수 있는 외부 링크를 제공하여, 외부 서버로부터 저작물을 제공한다. 웹툰 사이트는 웹툰 이미지 내 게시중인 저작권 침해사이트임을 알 수 있는 워터마크가 삽입되어 있다.

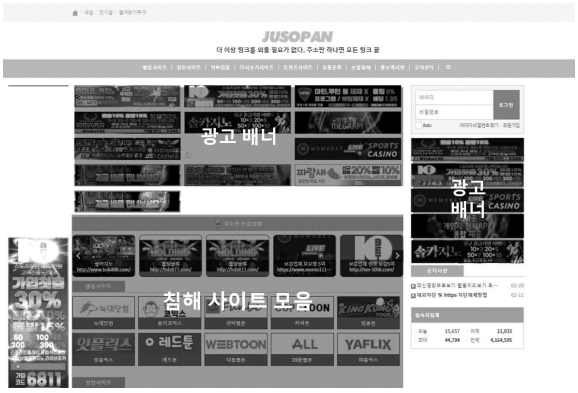
저작권 침해사이트에는 도박, 성인과 같은 정보가 게시된 불법 광고 배너가 공통적으로 존재하는 특징이 있다. 또한, 저작권 침해사이트에는 성인 및 도박 사이트로 직접 연결되는 링크가 존재하며, 저작권 침해사이트를 대표하는 이미지가 메인 페이지에 게시된 경우가 일반적이다[6]. <표 1>은 침해사이트와 광고 배너에서 높은 빈도수로 나타나는 키워드의 리스트를 나타내며, 제안기법의 사이트 침해 여부를 판단하는 기준으로 활용된다.

<표 1> 침해사이트 내 키워드 리스트

	Keyword List
토렌트	‘torrent’, ‘magnet’, ‘seed’ 등
영상스트리밍	‘HDVid’, ‘HLSPlay’ 등
웹툰	‘BL’, ‘GL’, ‘완결 웹툰’ 등
광고 배너	‘먹튀’, ‘매충’, ‘첫충’ 등

2.3 링크모음 사이트 특징

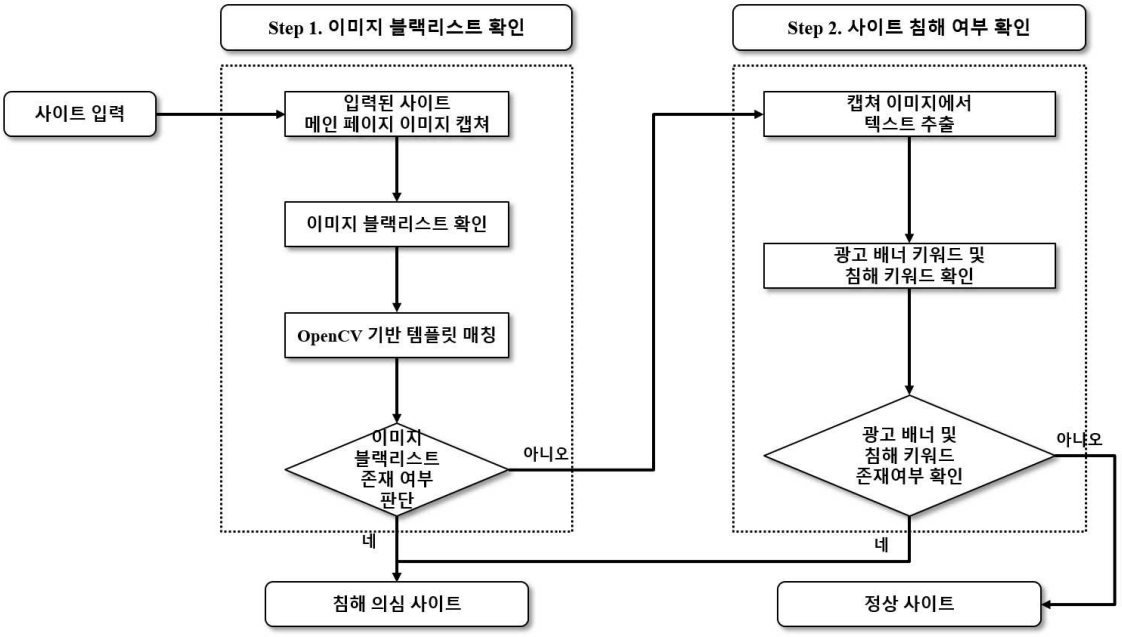
링크 모음 사이트는 저작권 침해사이트들을 리스트화하여 클라이언트에게 제공하는 사이트를 말한다. 링크 모음 사이트는 내부에 불법 광고 배너를 게시하여 수익을 창출하며, 다수의 저작권 침해사이트들이 배너 형태로 존재한다. 링크 모음 사이트의 배너는 저작권 침해사이트의 대표 이미지를 게시하여 사용자에게 가시적으로 알 수 있게 하는 경우가 일반적이며, 저작권 침해사이트의 메인 페이지에 이미지가 게시된다[6,7]. (그림 1)은 링크 모음 사이트의 특징을 나타낸다.



(그림 1) 링크 모음 사이트의 특징

3. 이미지 블랙리스트 데이터셋

본 장에서 제안하는 이미지 블랙리스트 데이터셋은 침해사이트 대표 이미지로 구성된다. 저작권 침해사이트는 일반적으로 자신의 웹사이트를 대표하는 이미지를 메인 페이지에 함께 게시하는



(그림 2) 블랙리스트 기반 저작권 침해 의심 사이트 탐지 기법

특징이 존재한다. 이를 이용하여, 입력된 웹사이트 내부에 블랙리스트 이미지 존재 여부를 통해 침해사이트 여부를 판단한다. 이미지 블랙리스트 데이터셋은 링크 모음 사이트 내부에 침해사이트별 대표 이미지가 존재하는 특징을 이용하여, 링크 모음 사이트 내부 이미지를 크롤링하여 구성한다.

4. 이미지 블랙리스트 기반 저작권 침해 의심 사이트 탐지 기법

본 장에서는 이미지 블랙리스트 기반 저작권 침해 의심 사이트 탐지 기법에 대하여 제안한다. (그림 2)는 본 장에서 제안하는 기법의 흐름을 나타낸다.

Step 1. 이미지 블랙리스트 확인

이 단계는 입력된 사이트의 메인 페이지 이미지와 이미지 블랙리스트 내부에 있는 저작권 침해사이트 대표 이미지 존재 여부를 판단하는 과정이다.

먼저, 입력된 사이트에 대하여 파이썬 Selenium 모듈을 이용하여 메인 페이지 스크린샷 이미지를 저장한다. 저장된 메인 페이지 이미지와 블랙리스트 이미지를 파이썬 OpenCV 모듈 내에 존재하는 템플릿 매칭을 이용하여, 메인 페이지 이미지 내 침해사이트 대표 이미지 존재 여부를 판단한다. 이때, 입력된 사이트 내부에 블랙리스트 이미지가 존재하면, 입력된 사이트를 침해 의심 사이트로 판단하고, 아닐 경우 Step 2를 분석한다.

Step 2. 사이트 침해 여부 확인

이 단계는 입력된 사이트의 침해 여부를 판단하는 과정으로, 광고 배너 키워드 리스트와 침해 키워드 리스트 존재 여부를 판단하는 과정이다.

먼저, 저장된 메인 페이지 이미지에 대하여, Google Vision API의 OCR 기능을 이용하여 이미지 내 텍스트를 추출한다. OCR 기능을 이용하여 광고 배너 내부에 존재하는 텍스트뿐만 아니라 메인 페이지 내 가시적으로 보이는 텍스트도 추출할 수 있다. 추출한 텍스트에 대하여 침해사이트별 저작권 침해 키워드와 광고 배너에서 사용되는 키워드 존재 여부를 검사한다. 침해 키워드 리스트는 침해사이트 내에서 높은 빈도수로 사용되는 키워드이다. 이때, 저작권 침해 키워드와 광고 배너 키워드가 추출된 텍스트 내에 존재하면 침해 의심 사이트로 판단하고, 아닐 경우 정상 사이트로 판단한다.

5. 결론

본 논문에서는 저작권 침해를 방지하기 위해 이미지 블랙리스트 기반의 저작권 침해 의심 사이트 탐지 기법을 제안하였다. 저작권 침해사이트의 메인 페이지 내부에 침해사이트를 대표하는 이미지가 존재하는 특징을 이용하여, 링크 모음 사이트에 존재하는 이미지를 블랙리스트로 정의하였다. 본 논문에서 제안한 기법을 통해 저작권 침해 의심 사이트 탐지 및 차단에 기여할 수 있을 것이다.

참고문헌

[1] 한국저작권보호원, “2018년 기준 불법복제물 유통실태 조사”, 저작권 보호 연차보고서, 2019, p.59

[2] 김정숙, 김태훈, “불법 링크사이트 실태조사연구”, 한국저작권위원회, 2018, pp.21-65

[3] 한국저작권보호원, “2016년 기준 불법복제물 유통실태조사”, 저작권 보호 연차보고서, 2017, p.25

[4] S. Choi and J.Kwak, “Feature Analysis and Detection Techniques for Piracy Sites”, KSII Transactions on Internet and Information Systems, Vol. 14, No. 5, pp. 2204-2220, 2020

[5] 강희광, 김창익, “딥러닝 기술을 이용한 템플릿 매칭 구현”, 대한전자공학회 학술대회, 2016, pp.343-345

[6] E. J Kim and J. Kwak, “Intelligent Piracy Site Detection Technique with High Accuracy”, KSII Transactions on Internet and Information Systems, Vol. 15, No. 1, pp.285-301, 2021

[7] E. J Kim, D.H Kim and J. Kwak, “Automated Detection Architecture for Piracyt Sties Based on Link Collection Sites”, The 12th International Conference on Internet(ICONI), 2020, pp.403-406,