이미지 메타데이터 기반 검증 및 콘텐츠 적절성 탐지 기법 연구

강영훈[1], 박세진[2]
[1]계명대학교 컴퓨터공학과 학부생
[2]계명대학교 컴퓨터공학과 교수
totohoon4@naver.com, baksejin@kmu.ac.kr

Metadata-Based Image Verification and NSFW Detection

Yeonghun Kang[1], Sejin Park[1]
[1]Department of Computer Engineering, Keimyung University

요 약

본 논문은 디지털 이미지의 신뢰성과 콘텐츠 적절성을 판단하기 위한 서버 기반 통합 검증 시스템을 제안한다. EXIF 메타데이터 검증, SHA-256 기반 위·변조 감지, NSFW 필터링 기능을 통합하여, 이미지 업로드 시 서버가 자동 분석을 수행하고 검증된 콘텐츠만 저장되도록 구성하였다.

NSFW 분류 실험은 DarkyMan/nsfw-image-classification 데이터셋 기반으로, 다양한 시각적 특성을 반영한 1805장의 이미지로 수행되었으며, ViT 기반 Falconsai 모델이 62.88%의 정확도를 기록하였다. 이는 fine-tuning 없이 사전학습 모델을 적용한 결과로, 구조적 적용 가능성과 개선 여지를 평가하는 데 의의가 있다. 또한 본 시스템은 WebSocket 기반 위치 연동, RDS 및 S3 저장소 구성, 인증 기반 접근 제어 등 보안 요소 확장 가능성도 함께 고려하였다.

1. 서론

디지털 이미지에는 촬영 위치, 시간, 기기 정보 등이 담긴 EXIF 메타데이터가 포함되나, 이는 쉽게 조작될 수 있어 위치 기반 플랫폼에서 이미지의 신뢰성에 문제가 발생할 수 있다.

본 논문은 EXIF 검증, SHA-256, NSFW 필터링을 통합해 이미지의 정합성과 적절성을 자동 분석하고 신뢰도를 높이고자 한다.

2. 관련 연구

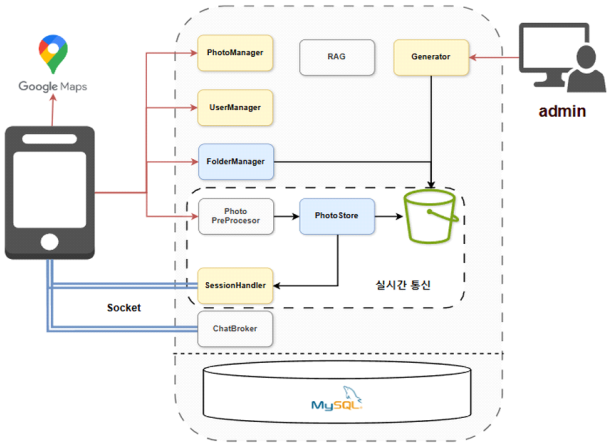
이미지 메타데이터 기반 검증 기법은 GPS 정보 조작 탐지, 메타데이터 유무 분석, 표준화 구조(MPEG-7 등) 활용 방식으로 발전해 왔다[1-3]. 또한 픽셀 단위 분석을 통한 위변조 탐지 기법도 함께 제안되었다[4].

그러나 대부분의 기존 연구는 메타데이터 분석에만 집중되어 있으며, 이미지 콘텐츠 자체에 대한 분석은 미흡하다는 한계가 있다. 본 논문은 이에 대응하여, SHA-256 기반 위·변조 감지와 NSFW 분류 모델을 통합한 서버 기반 사전 검증 시스템을 제안한다.

3. 제안 기법

본 논문은 EXIF 메타데이터 검증, SHA-256 해시 기반 위·변조 감지, NSFW 콘텐츠 필터링을 통합한 서버 기반 이미지 사전 검증 시스템을 제안한다. 서버는 업로드된 이미지를 자동 분석한 후, 검증 결과에 따라 저장 여부를 판단한다.

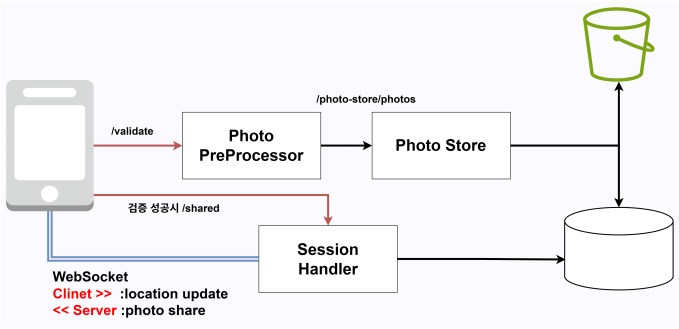
전체 시스템 구성은 그림 1과 그림 2에 나타낸다.



(그림 1) 이미지 콘텐츠 검증 시스템 아키텍처

그림 1은 제안하는 시스템의 전체 처리 구조를 나

타낸 것으로, 클라이언트와 서버 간 데이터 흐름 및 저장 경로를 포함한다.



(그림 2) 서버 기반 이미지 검증 및 저장 처리 흐름

그림 2는 이미지 업로드 시 서버가 수행하는 검증 처리 과정을 나타낸다.

사용자가 이미지를 업로드하면, 서버는 메타데이터 분석과 NSFW 필터링을 수행한 후, 공유 설정과 프레임 적용 여부에 따라 저장 경로를 분기한다. 공유 설정이 활성화된 경우, 해당 이미지는 RDS와 S3 버킷에 저장되며, 위치 정보는 WebSocket을 통해 실시간으로 Session Handler에 전달된다.

제안 시스템은 다음 항목을 기준으로 이미지의 신뢰성과 적절성을 평가한다.

1. 촬영 위치 검증: EXIF GPS 정보와 업로드 당시 기기 위치를 비교하여 조작 여부를 분석한다.

2. 시간 무결성 확인: 타임스탬프와 서버 수신 시간 차이를 통해 시간 위조 여부를 확인한다.

3. 기기 정보 일관성 분석: 동일 사용자 계정 내 이미지 간 기기 정보(카메라 모델, 해상도 등)의 일치 여부를 비교한다.

4. SHA-256 해시 적용: 이미지의 위·변조 여부를 확인하며, 향후 무결성 검증 기능으로의 확장 가능성을 가진다. 서버는 업로드 시 해시를 생성하며, 이후 비교를 통해 변조 여부를 검증한다.

5. Not Safe For Work (NSFW) 필터링: 사전학습된 분류 모델 세 가지(Falconsai, LukeJacob, AdamCodd)의 성능을 비교하였다. 실험은 DarkyMan/nsfw-image-classification 데이터셋을 기반으로 하되, 다양한 시각적 특성을 포함한 복합 이미지 샘플을 구성하여 수행되었다. 전체 1805장의 이미지(N: NSFW 722장, SFW 1083장)를 활용하였으며, ViT(Visual Transformer) 기반 Falconsai 모델이 62.88%로 가장 높은 분류 정확도를 기록하였다. 모델별 성능 비교 결과는 <표 1>에 정리하였다.

<표 1> NSFW 분류 모델 성능 비교

모델명	분류 정확도 (%)	모델 구조	백본 세부
Falconsai	62.88	Transformer 기반 (ViT)	ViT-B/16
LukeJacob	60.78	CNN 기반	ResNet-50
AdamCodd	61.22	ViT-Base (Hugging Face)	ViT-B/16

4. 결론

본 논문은 EXIF 메타데이터 검증, SHA-256 기반 위·변조 감지, NSFW 필터링을 통합한 서버 기반 이미지 검증 시스템을 제안하였다. 제안 구조는 촬영 정보의 정합성과 콘텐츠 적절성을 자동 분석하도록 구성되어 있으며, 이미지의 신뢰성과 안전성 향상이 기대된다.

NSFW 분류 실험에서는 fine-tuning 없이 사전학습된 Falconsai 모델이 62.88%의 가장 높은 정확도를 기록하였다. 이는 구조적 적용 가능성과 향후 성능 개선 여지를 확인하는 데 의의가 있으며, 해당 수치는 절대적인 판단보다는 실용 가능성과 확장성 중심으로 해석할 수 있다.

본 시스템은 이미지 업로드 시 메타데이터와 콘텐츠 정보를 자동 분석한 후, 검증된 콘텐츠만 저장되도록 설계되었으며, 저장 및 전송 과정에서의 보안 강화를 위한 기술 적용 가능성도 함께 고려하였다. 예를 들어, SHA-256 해시 알고리즘을 활용한 무결성 검증, WebSocket 기반 위치 연동, 인증 기반의 저장소 접근 제어 등이 포함되며, 향후 사용자 인증, 접근 제어, 전송 계층 보안(TLS) 등의 요소를 반영한 보안 체계 확장이 기대된다.

참고문헌

[1] 오경수, "사진 메타데이터의 처리방법에 대한 연구", 정보과학회논문지, 2012.

[2] 이창수 외, "다중 메타데이터 기반 JPEG 이미지 파일 출처 식별: 모바일 메신저 앱 16종을 대상으로", 디지털포렌식학회논문지, 2021.

[3] 전인배, "이미지 및 비디오 메타데이터에 관한 연구", 한국멀티미디어학회논문지, 2009.

[4] Bianchi, T., Piva, A., & Barni, M., "Image forgery localization via block-grained analysis and classification," International Journal of Computer Vision, 2011.