# 실시간 유튜브 혐오 썸네일 필터링 웹서비스 개발

김찬O, 박상근

경희대학교 소프트웨어융합학과

rlacks2593@khu.ac.kr, sk.park@khu.ac.kr

# Development of a Real-Time Web Service For Filtering Disgusting Youtube Thumbnails

Chan KimO, Sangkeun Park

Department of Software Convergence, Kyung Hee University

**요 약**

본 연구에서는 동영상 플랫폼에서 부적절한 썸네일로 인한 사용자 불쾌감을 줄이기 위해 혐오 이미지를 실시간으로 필터링하는 알고리즘 및 웹서비스를 제안한다. 기존 연구는 폭력적이거나 선정적인 콘텐츠 필터링에 초점을 맞췄으나, 징그러운 벌레나 모공 등의 혐오 이미지 필터링은 상대적으로 덜 주목받았다. 이에 혐오 이미지를 정의하고 유튜브 접속 시 실시간 혐오 썸네일 필터링이 크롬 플러그인으로 구현하였다. 사용자 스터디를 통해 알고리즘의 사용성을 검증한 결과, 본 연구에서 제안한 모델이 유튜브 사용 시 혐오 이미지로 인한 부정적 감정을 줄이는 데 효과적일 수 있음을 확인했다

.

## 1. 서 론

유튜브 등의 동영상 플랫폼은 사용자가 여러 동영상을 한눈에 파악할 수 있도록 영상마다 썸네일(미리보기 이미지)을 제공하고 있다. 이 영상 썸네일은 사용자의 시청 만족도와 재시청 의도에 유의미한 영향을 준다[1]. 썸네일은 해당 영상에서 다루는 콘텐츠를 빠르게 파악할 수 있다는 장점이 있지만, 선정적인 장면, 폭력적인 장면 등 썸네일 자체에 부적절한 정보가 포함되어 있다면 사용자에게 불쾌감을 줄 수도 있다 [2].

썸네일은 콘텐츠 제작자가 영상을 등록할 때 직접 지정하거나 동영상 플랫폼에서 제공하는 알고리즘에 의해 자동으로 선택되므로, 동영상 플랫폼 사용자가 예기치 못한 썸네일을 갑작스럽게 보게 되었을 때 부정적인 경험을 할 수도 있다. 특히, 혐오 이미지가 썸네일로 사용된다면, 사용자 뇌의 편도체를 활성화해 부정적인 감정 및 정서적 불안감을 촉진하고 정서를 해칠 위험이 있다 [3,4].

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2024년도 SW중심대학사업의 결과로 수행되었음"(2023-0-00042)

영상 또는 이미지에서 폭력적이거나 선정적인 컨텐츠를 필터링하는 다양한 연구가 수행되고 있다. 유해 객체를 감지하고 필터링하거나[5], 장면의 색 분포를 기반으로 선정적인 영상을 분류하는 연구[6]가 시도되었다. 하지만 선정적이거나 폭력적인 이미지 또는 영상 외에 법적으로 규제를 받지 않는 컨텐츠(예: 징그러운 벌레, 모공 등) 필터링에 관한 연구는 상대적으로 덜 주목받고 있다. 본 연구에서는 법적 제재를 받지 않는 혐오 이미지를 정의하고, 이를 실시간으로 필터링하는 알고리즘을 개발하였다. 이 알고리즘을 유튜브 영상 시청 시 사용할 수 있도록 크롬 플러그인을 개발했으며, 사용자 스터디를 통해 해당 알고리즘의 사용성을 검증했다.

## 2. 관련 연구

미디어에서 사용자에게 부정적인 경험을 제공할 수 있는 이미지 또는 영상을 필터링하기 위한 다양한 연구가 수행되었다. Maheswaran et al.[5]은 이미지에 존재하는 폭력 관련 객체를 분석하고 필터링했으며, Tofa et al.[6]은 색상 분포를 활용해서 영상에서 선정적인 장면을 분류했다. Tahir et al.[7]은 영상 분할

프레임과 음성, 영상 내 객체 움직임을 함께 사용해 폭력 및 선정성 컨텐츠를 분류했으며, Yekbote et al.[8]은 비디오 ID, 제목, 컨텐츠 유형과 같은 메타데이터와 영상을 접목해 폭력과 선정적 영상을 분류했다. 기존 연구는 이러한 폭력성 및 선정성과 같이, 각 나라 및 플랫폼마다 규정을 갖고 관리하는 콘텐츠 필터링에 집중하고 있다. 이에 따라, 징그러운 벌레, 모공 등과 같이, 법적으로 문제가 되지는 않을 수 있지만 시청자로 하여금 혐오감을 느낄 수 있을 법한 이미지 및 영상 관련 필터링 연구는 상대적으로 덜 주목을 받고 있다.

혐오감에 관한 기준은 개인마다 크게 다를 수 있어서 필터링하기 위한 명확한 기준을 잡기는 쉽지 않다. Oh et al.[9]은 온라인 설문 조사를 통해 혈흔, 벌레, 환공포증, 더러움, 네 가지를 시각적 혐오감을 일으킬 수 있는 범주로 정리했다. Cray et al.[10]은 혐오에 관한 기존 연구를 기반으로 상한 음식, 질병 관련 자극 및 배설물을 혐오감 자극으로 선정했으며, Mataix-Cols et al.[11]은 혐오 영상을 이용한 실험을 통해 일반인이 훼손된 신체, 상처, 벌레에 혐오감을 느낄 수 있다는 사실을 확인했다

본 연구에서는 이미 일반적인 규제 대상에 속하는 폭력성 선정성 외에, 법적 규제를 위배하지 않으면서 사용자에게 시각적 혐오감을 유발할 수 있는 요소로 1) 심한 피부질환, 2) 많은 벌레, 3) 동물 사체, 4) 환공포증으로 정했다. 이를 기반으로, 시각적 혐오감을 일으킬 수 있는 요소가 포함된 썸네일을 실시간으로 필터링하기 위한 서비스를 제안한다.

### 3. 혐오 썸네일 분류 모델 개발

혐오감을 일으키는 이미지인지 판단하기 위한 모델을 개발하기 위해, 파이썬 Selenium 모듈을 이용하여 유튜브에서 썸네일 이미지를 수집했다. 우선, 유튜브의 혐오 관련 채널 및 검색을 통해 1) 심한 피부질환, 2) 많은 벌레, 3) 동물 사체, 4) 환공포증, 총 4개 카테고리의 혐오 썸네일 이미지를 크롤링했다. 각 범주당 600개씩 총 2,400개의 혐오 썸네일 이미지를 수집했다. 연구진이 직접 눈으로 확인하면서 혐오와 관련 없는 썸네일 이미지는 제거하고 다른 혐오 썸네일로 대체했다. 혐오 썸네일이 아닌 정상 썸네일을 수집하기 위해 예능, 음악, 스포츠, 게임, 인터넷 방송 등의 인기 동영상에서 혐오와 무관한 정상 썸네일 이미지 600장을 수집했다.

ResNet50 모델의 입력 크기(224*224)에 맞게, 모든 썸네일 이미지의 크기를 변경했다. 카테고리별로 500장씩은 트레이닝 데이터로, 100장씩은 테스트 데이터로 분류했다. 학습 모델은 실시간 이미지 분류에 적합한 경량화 모델 중 지연시간이 짧고 성능이 우수한 ResNet50[12]을 학습 모델로 선정했다. 데이터 증강을 위해, 파이썬 Keras[2] 패키지의 ImageDataGenerator를 이용했다. 트레이닝 데이터에 회전, 평행이동, 확대 등을 적용하고, 각 카테고리의 트레이닝 데이터를 500개에서 1000개로 증강했다. 이 트레이닝 데이터를 활용해 모델 학습을 진행했다. 테스트 데이터에 대해 모델 Optimizer를 Adam, Adagrad, RMSprop을 사용하고, 테스트한 결과는 [표1]과 같다. [표1]의 결과를 바탕으로, 본 연구에서 시스템에 사용할 최종 모델로 Adam을 선정했다.

표1. 혐오 썸네일 필터링 모델의 성능 평가 결과

| Optimizer | Accuracy | F1-score | Precision | Recall |
|---|---|---|---|---|
| Adam | 0.79 | 0.79 | 0.83 | 0.75 |
| Adagrad | 0.70 | 0.73 | 0.75 | 0.71 |
| RMSprop | 0.78 | 0.77 | 0.81 | 0.73 |

### 4. 혐오 썸네일 분류 모델 개발

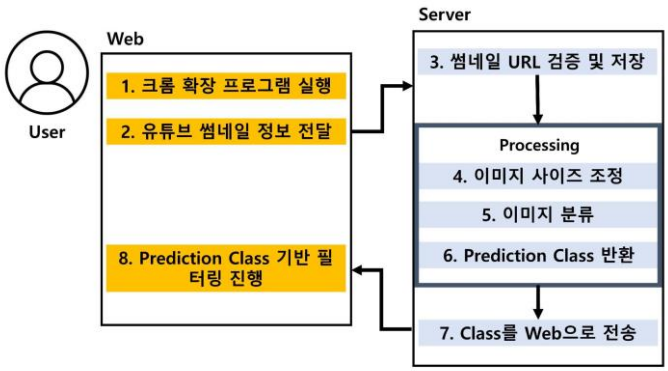


그림 1. 혐오 썸네일 필터링 서비스 구조도

학습된 혐오 이미지 분류 모델을 활용해서, 사용자가 유튜브에 접속하면 혐오 썸네일을 실시간으로 필터링해 주는 크롬 플러그인 익스텐션 서비스[3]를 개발했다. 해당 서비스의 구조는 [그림 1]과 같다. JavaScript 기반의 크롬 익스텐션 앱을 개발하고, Python 웹 프레임워크인 Flask를 이용해 서버를 구축했다. 사용자가 유튜브에 접속하면 화면에 뜨는 모든 썸네일 이미지 URL이 서버로 전달되고, 서버는 전달받은 썸네일 이미지 URL을 통해 썸네일 이미지를 다운로드하고 이미지 크기를 조정한다. 학습시킨 ResNet50 모델을 통해 각 썸네일 이미지는 1) 심한 피부질환, 2) 많은 벌레, 3) 동물 사체, 4) 환공포증, 5) 일반 중 하나로 예측된다. 예측 결과는 다시 크롬 익스텐션에 전달된다. 크롬 익스텐션은 예측 결과가 혐오 카테고리 중 하나로 분류되면,

[2] https://keras.io/api/

[3] https://tinyurl.com/36db65bn

사용자가 혐오 이미지를 바로 볼 수 없도록 혐오 이미지로 분류된 썸네일을 별도로 준비된 Blocked 이미지로 교체한다 [그림2, 3].

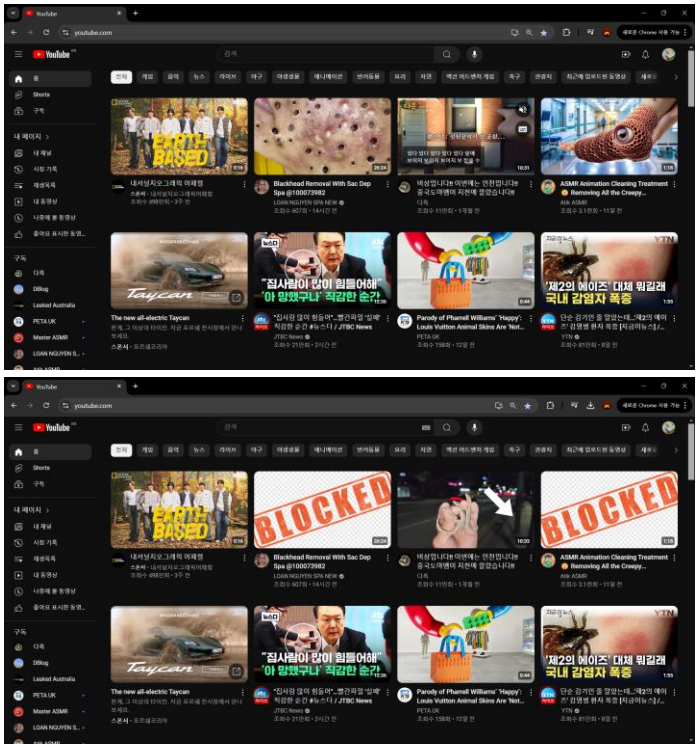


그림 2,3. 유튜브 혐오 썸네일 필터링 앱 적용 전/후

## 5. 사용자 스터디

2024년 10월, 서비스의 사용성을 검증하기 위해 평소 유튜브를 즐겨보는 20대 남자 2명, 50대 여자 1명을 대상으로 사용자 스터디를 진행했다. 유튜브에서 혐오 썸네일을 목격한 경험에 대한 질문에, 20대 남성 2명이 실제로 심한 피부질환과 관련된 혐오 썸네일을 목격한 경험이 있다고 응답했다. 참여자는 혐오 썸네일 필터링 서비스에 대한 설명을 듣고, 약 30분 동안 유튜브 썸네일을 자유롭게 탐색해달라고 요청했다. 사용자는 평소와 같이 유튜브를 사용하고, 혐오 썸네일이 등장할 때 자동으로 필터링 기능을 사용했다. 서비스의 필요성을 검증하고자 혐오 썸네일 필터링의 필요성을 질문하고, 5점 척도 기반 조사를 통해 평균 4.17과 표준편차 0.47의 결과를 얻었다. 20대 남 1은 “의도치 않게 혐오 썸네일을 본 적이 있는데, 미리 분석을 통해 필터링해 줘서 좋았다”라는 긍정적인 반응을 보였다. 20대 남 2는 “필터링할 때 대체 이미지의 변경이나 모자이크 등을 고려했으면 좋겠다.”라는 피드백을 제시했다.

## 6. 결론 및 제언

본 연구에서는 시각적으로 혐오감을 주는 썸네일을 필터링하기 위해 혐오 썸네일을 분류하는 모델을 개발하고, 이를 기반으로 유튜브에서 실시간으로 혐오 썸네일을 필터링할 수 있는 크롬 플러그인을 개발했다. 본 연구를 확장하면 사용자 개인의 요구에 맞는 카테고리에 대해 맞춤 필터링을 제공할 수 있고, 이는 사용자의 만족도를 높이고 사용성을 증가하는 데 도움이 될 수 있다. 하지만 공개적인 자료가 적은 데이터의 특성과 짧은 수집 기간으로 정확도가 떨어진다는 한계가 존재한다. 향후 연구에서는 사용자 설문조사 등을 이용하여 시각적 혐오를 재정의 및 확장하고, 충분한 데이터를 수집하여 모델의 정확성과 서비스의 활용성을 높이고자 한다. 또한, 필터링 방식을 세분화하여 사용자의 만족도를 더 높일 계획이다.

## 7. 참고문헌

[1] 이승민, “유튜브 섬네일의 시각 표현 요소가 사용자 만족도와 재시청의도에 미치는 영향 연구”, 디지털컨텐츠학회논문지, 948-949, 2022

[2] 편슬기, “유튜브, 혐오·자극적 썸네일 “역겹고 불편해”이용자 불만 폭발”, 시사 오늘, 2023.08.14., https://tinyurl.com/4k8sjm96

[3] Schienle et al, “Brain activation of spider phobics towards disorder-relevant, generally disgust- and fear-inducing pictures”, Neuroscience Letters Volume 388, 1-6, 2005

[4] 강은영, “부정적인 당신, 긍정의 힘을 얻고 싶다면? 편도체의 흥분을 억제해라!”, 한국강사신문, 2021.10.08., https://tinyurl.com/yh959m4h

[5] Maheswaran et al. “YOLO based Efficient Vigorous Scene Detection And Blurring for Harmful Content Management to Avoid Children’s Destruction”, ICESC, 2023

[6] Tofa et al. “Inappropriate scene detection in a video stream”, Brac University, 2017

[7] Tahir et al. “Bringing the Kid back into Youtube Kids:Detecting Inappropriate Content on VideoStreaming Platforms”, 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019

[8] Yekbote et al. “A Novel Approach for Inappropriate Content Detection and Classification of Youtube Videos using Deep Learning”, ICAAIC, 2024

[9] Oh et al. “Development and utilization of a disgusting image dataset to understand and predict visual disgust”, Image and Vision Computing Volume 72, 24-38, 2018

[10] Cray et al. “The sensory channel of presentation alters subjective ratings and autonomic responses toward disgusting stimuli—Blood pressure, heart rate and skin conductance in response to visual, auditory, haptic and olfactory presented disgusting stimuli”, Frontier in Human Neuroscience Volume 7, 2013

[11] Mataix-Cols et al. “Individual differences in disgust sensitivity modulate neural responses to aversive/disgusting stimuli”, European Journal of Neuroscience: Volume 27, Issue 11., 3050-3058, 2008

[12] He et al. “Deep Residual Learning for Image Recognition”, CVPR, 2016