JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

기반 기술

J. Adv. Navig. Technol. 25(6): 557-561, Dec. 2021

영상 콘텐츠의 오디오 분석을 통한 메타데이터 자동 생성 방법

Method of Automatically Generating Metadata through Audio Analysis of Video Content

용성중 · 박효경 · 유연휘 · 문일영*
한국기술교육대학교 컴퓨터공학과

Sung-Jung Young· Hyo-Gyeong Park · Yeon-Hwi You· Il-Young Moon*

Department of Computer Science and Engineering, Korea University of Technology and Education, Cheonan, 31253, Korea

[요 약]

영상 콘텐츠를 사용자에게 추천하기 위해서는 메타데이터가 필수적인 요소로 자리 잡고 있다. 하지만 이러한 메타데이터는 영상 콘텐츠 제공자에 의해 수동적으로 생성되고 있다. 본 논문에서는 기존 수동으로 직접 메타데이터를 입력하는 방식에서 자동으로 메타데이터를 생성하는 방법을 연구하였다. 기존 연구에서 감정 태그를 추출하는 방법에 추가로 영화 오디오를 통한 장르와 제작국가에 대한 메타데이터 자동 생성 방법에 대해 연구를 진행하였다. 전이학습 모델인 ResNet34 인공 신경망 모델을 이용하여 오디오의 스펙트로그램으로부터 장르를 추출하고, 영화 속 화자의 음성을 음성인식을 통해 언어를 감지하였다. 이를 통해 메타데이터를 생성 인공지능을 통해 자동 생성 가능성을 확인할 수 있었다.

[Abstract]

A meatadata has become an essential element in order to recommend video content to users. However, it is passively generated by video content providers. In the paper, a method for automatically generating metadata was studied in the existing manual metadata input method. In addition to the method of extracting emotion tags in the previous study, a study was conducted on a method for automatically generating metadata for genre and country of production through movie audio. The genre was extracted from the audio spectrogram using the ResNet34 artificial neural network model, a transfer learning model, and the language of the speaker in the movie was detected through speech recognition. Through this, it was possible to confirm the possibility of automatically generating metadata through artificial intelligence.

Key word : Audio, AI, Metadata, Recommendation System, Voice Recognition.

https://doi.org/10.12673/jant.2021.25.6.557

Received 30 November2021; **Revised** 1 December 2021
Accepted (Publication) 26 December 2021 (30 December 2021)

***Corresponding Author ; Il-Young Moon**

Tel: +82-041-560-1493
E-mail: iymoon@koreatech.ac.kr

Ⅰ. 서 론

추천 시스템은 사용자와 아이템의 상호작용을 모델링 하여 개인화된 경험을 제공하는 것을 목표로 한다. 새롭게 생성되는 정보와 물품의 양이 점점 늘어나면서, 사용자의 합리적인 선택을 도와주는 보조적인 역할을 수행하고 있다. 쇼핑, 음악 및 영화 스트리밍 등의 서비스를 통해 우리 삶 속 깊숙하게 들어와 있으며, 의료, 정치, 교통 등의 다양한 영역에도 점차 활용되고 있다 [1].

최근 IPTV(internet protocol television)와 스마트 TV의 등장과 영화나 TV 프로그램 같은 영상 콘텐츠를 검색하고 시청할 수 있는 웹 서비스의 등장으로 사용자는 영상 콘텐츠에 접근이 용이해졌다. 기존의 영상 콘텐츠 시청 방법과 달리 시간에 구애받지 않고 상영이 되었던 영화나 TV 프로그램의 검색과 시청을 할 수 있게 되었다. 이에 따라 원하는 영상 콘텐츠를 찾고자 하는 욕구가 증가하였다 [2].

디지털 시스템에서 콘텐츠를 데이터로 저장하고 저장된 데이터를 이용자에게 필요한 콘텐츠로 표현하고 기술하는 역할과 기능을 메타데이터가 맡고 있다 [3].

즉, 영상 콘텐츠를 사용자에게 추천하기 위해서는 메타데이터가 필수적인 요소로 자리 잡고 있다. 하지만 이러한 메타데이터는 영상 콘텐츠 제공자에 의해 수동적으로 생성된다.

기존 연구에서는 미디어에서 추출한 오디오를 통하여 감정에 대한 태그를 인공지능을 통해 자동으로 추출하였다 [4]. 본 논문에서는 영상 콘텐츠의 오디오 분석을 통해 장르와 제작국가에 대한 메타데이터 자동 생성 방법에 대해 추가 연구를 진행하였다.

Ⅱ. 연구방법

2-1 연구방법 설정

기존 연구에서는 미디어에서 추출한 오디오를 통하여 감정에 대한 태그를 추출하는 연구를 진행하였다.

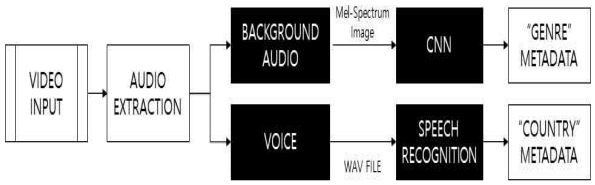

그림 1. 오디오 추출 및 메타데이터 생성 알고리즘
Fig. 1. Audio Extraction and Metadata Generation Algorithms

감정에 대한 태그는 MFCC(mel-frequency cepstral coefficient)를 활용하여 음성의 특징을 추출하고 SVM(support vector machine) 기계학습으로 감정에 대한 특정 카테고리로 분리하였다 [4]. 본 연구는 기존 연구에서 감정 태그를 추출하는 방법에 추가로 영화의 오디오를 통해 장르와 제작국가 메타데이터 생성을 위한 연구를 진행하였다. 그림 1과 같이 영화 예고편을 사용하여, 영상에서 오디오를 추출하여 배경 오디오와 음성으로 각각 분리하였다. 배경 오디오는 STFT(short time fourier transform)를 활용하여 Mel-Spectrogram 이미지를 획득하였고, 획득한 Mel-Spectrogram 이미지를 CNN(convolutional neural network)에 적용하였다. 신경망에 “액션, 코믹, 공포, 멜로” 4가지 장르를 인식하는 딥러닝 모델을 생성하였다. 음성 오디오는 음성인식(Speech Recognition)을 통해 영화에서 화자의 말을 문자로 변환하여 변환된 문자가 “한국어, 영어, 일어, 중국어”로 구분하여 국가명을 생성하였다.

2-2 장르 메타데이터 생성 연구

오디오 신호 처리를 위해 시간에 따른 주파수 성분의 변화에 대해 분석하고자 STFT 분석 기법을 사용하였다. STFT는 일반적으로 사용하는 FFT(fast fourier transform)보다 시간-주파수 영역을 모두 분석할 수 있어 그림 2, 그림 3과 같이 장르별 배경음악 스펙트로그램 이미지를 생성하였다.

영화 콘텐츠에 대한 스펙트로그램을 획득하여 인공신경망 모델을 생성하기에는 학습 데이터 전처리 과정의 한계가 있어 전이학습을 진행하여 VGG-19구조를 뼈대로 컨볼루션 층을 추가하여 총 34개의 층으로 정확도를 높인 ResNet34 인공신경망 모델을 이용하여 진행하였다.

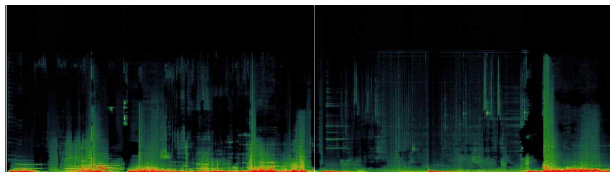

그림 2. 공포 영화 예고편 배경 오디오 스펙트로그램
Fig. 2. Spectrogram of Horror Movie Trailer Background Audio

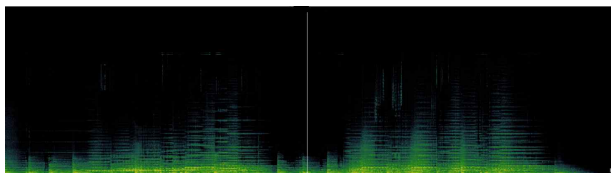

그림 3. 멜로 영화 예고편 배경 오디오 스펙트로그램
Fig. 3. Spectrogram of Melodrama Trailer Background Audio

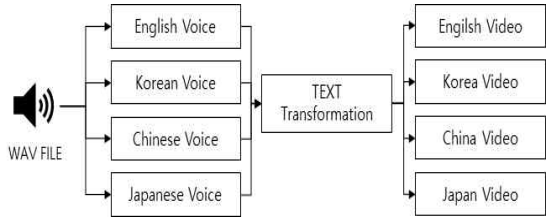

그림 4. 음성인식 처리 실험
Fig. 4. Speech Recognition Processing Experiment

전이학습은 학습 데이터가 적은 경우, 높은 성능을 확보하기 위해 많이 사용된다. 전이학습은 ImageNet이나 OpenImage와 같은 대용량 데이터 셋으로 사전 학습된 가중치를 사용하여 새로운 데이터 특징을 추출하고 분류하는 방법이다. 모든 층을 학습하지 않고 부분적인 학습을 진행하기 때문에 학습 시간 절감뿐만 아니라 성능에서도 효율적이다 [5]. ImageNet과 OpenImage는 대표적 범용 대규모 이미지 데이터 셋을 말하며, 이미지 분류 모델의 성능 평가를 위한 표준 평가데이터로 사용된다.

ResNet34는 VGG-19 구조를 바탕으로 컨볼루션 층을 추가하여 총 34개의 층으로 정확도를 높인 신경망 모델이다. VGGNet이 대표적인 일반적인 신경망은 층이 깊어질수록 성능이 뛰어나지만, 특정 지점 이상으로 층이 깊어진다면 기울기(Gradient) 소실(vanishing)과 폭발(exploding)문제가 발생해 성능이 떨어지는 결과를 나타낸다. 이러한 문제점을 개선하기 위해 skip/shortcut connection 더하였으며, 이에 따라 기울기 소실 문제를 해결하며 정확도가 감소하지 않고 신경망의 층을 깊게 쌓으며 더 나은 성능의 신경망을 구축할 수 있다 [6]. 또한, ResNet에서 사용하는 Bottleneck desgin은 신경망의 복잡도를 감소시키며, 기존 VGG-16 등과 같은 CNN 모델보다 처리 속도가 빠르기 때문에 해당 모델을 선택하게 되었다.

2-3 국가 메타데이터 생성 연구

영화 콘텐츠는 출연 인물들이 자국어를 사용하기 때문에 화자의 음성을 텍스트로 변환하고, 텍스트의 언어를 구분하여 국가에 관한 메타데이터를 획득하였다. 여기서 화자의 음성을 텍스트로 변환하기 위해 음성인식(STS; speech-to-text system) 기능을 활용하였다 [7]. 그림 4와 같이 구글의 "Speech Recognition" 라이브러리를 사용하여 추출된 음성 파일을 텍스트로 변환하여 저장 후 파이썬의 "Langdetect" 라이브러리를 통해 저장된 텍스트의 언어를 확인함으로써 국가를 확인할 수 있었다.

Ⅲ. 연구결과

장르별 배경음악을 "action, comic, mello, horror" 클래스로 분류하였으며, 학습 데이터 12개와 테스트 데이터를 장르별 1개씩 준비하여 Resnet34 이미지 인식 신경망을 사용하여 학습과 테스트를 진행하였다.

다음의 그림 5는 학습 데이터의 장르별 배치 결과이며, 그림 6은 장르별 테스트 스펙토그램이다. 학습 데이터 수가 적은 상황과 빠른 학습 속도를 고려하여 전이학습을 통해 신경망 특징 추출 능력을 그대로 사용하고 준비된 학습 클래스의 개수 만큼 모델의 출력 뉴런 수 4개를 교체 후 마지막 레이어에 다시 학습하는 방식으로 진행하였다. 학습은 총 50번의 Epoch을 진행하였고, 학습 후 테스트를 진행하여 표 1과 같이 예측에 대한 ACC와 LOSS의 평가 결과를 확인하였다.

이처럼 평가율이 높게 나타나고 있다는 것은 배경음악 스펙토그램을 바탕으로 장르 구분을 위한 학습 데이터를 제대로 인식하고 분류한다는 것을 확인할 수 있었다.

또한, 분리한 음성 오디오는 "Speech Recognition" 라이브러리를 통해 그림 7과 같이 언어를 감지하고 해당 언어를 문자로 출력하여 실제 내용과 일치도를 확인하였다. Fig. 7은 Recognizer 객체를 생성 후 음성 오디오 WAV 파일을 읽어 구글 Recognizer 객체를 통해 변수에 Result와 같이 문자로 저장된 것을 확인할 수 있다.

인식된 문자는 "Langdetect" 언어 감지 라이브러리를 통해 그림 8과 같이 detect 함수에 Fig. 7에서 변수에 저장된 문자를 전달하여 해당 언어를 감지하여 조건문을 통해 Result와 같이 국가를 출력하는 결과를 확인하였다.

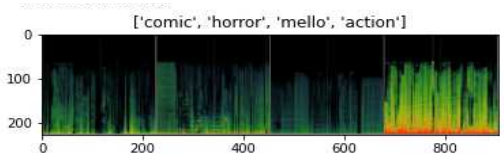

그림 5. 학습 데이터 장르별 배치 결과
Fig. 5. Result for Training Data Placement

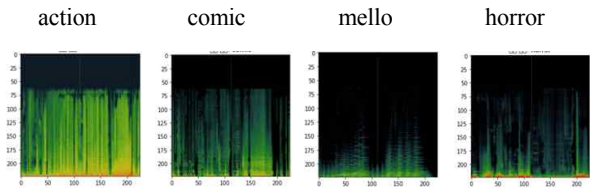

그림 6. 테스트 데이터 이미지
Fig. 6. Test Data Image

표 1. 테스트 데이터 예측 및 평가 결과

Table 1. Test Data Prediction and Evaluation results

Train Class		Test Data		Evaluation		
				Prediction	Loss	Acc
action	Godzilla Ashfall Transformers	action	Iron Mask	action	0.24	100%
comic	Extreme Job Wonderful Nightmare Honest Candidate	comic	Secret Zoo	comic	0.24	100%
mello	Noting Hill The Beauty Inside Once	mello	Carol	mello	0.27	100%
horror	0.0MHz Gonjiam Us	horror	The Nun	horror	0.22	100%

```
form langdetect import detect
import speech_recognition as sr

r = sr.Recognizer()
test_voice = sr.AudioFile('./test.wav')
with test_voice as source:
   voice = r.record(source)

str = r.recognize_google(voice)
print(str)
```

Result : the ministers are not to make their decision

그림 7. 구글 Colab에서의 음성 인식 실행 결과
Fig. 7. Voice Recognition Execution Result in Google Colab

```
form langdetect import detect
import speech_recognition as sr

r = sr.Recognizer()
test_voice = sr.AudioFile('./test.wav')
with test_voice as source:
   voice = r.record(source)

str = r.recognize_google(voice)

str_de = detect(str)
if str1 == "ko":
  print("input language Korean")
elif str1 == "en":
  print("input language English")
elif str1 == "zn-ch":
  print("input language Chinese")
if str1 == "ja":
  print("input language Japanese")
```

Result : input language English

그림 8. 구글 Colab에서의 언어 확인 실행 결과
Fig. 8. Language Check Execution Result in Google Colab

이러한 학습 결과는 영상 콘텐츠에서 추출한 배경음악을 이용하여 인공지능이 장르를 구분하고 언어 감지를 통하여 장르 및 해당 국가에 대한 메타데이터 자동 생성 방법에 사용할 수 있을 것이다.

Ⅳ. 결 론

디지털 시스템에서 콘텐츠를 데이터로 저장하고 저장된 데이터를 이용자에게 필요한 콘텐츠로 표현하고 기술하는 역할과 기능을 메타데이터가 맡고 있으며, 영상 콘텐츠 또한 사용자에게 추천하기 위해서는 메타데이터가 필수적인 요소로 자리잡고 있다. 하지만 이러한 메타데이터는 영상 콘텐츠 제공자에 의해 수동적으로 생성되어 본 논문에서는 기존 수동으로 직접 메타데이터를 입력하는 방식에서 영상 콘텐츠의 장르, 제작 국가에 대한 메타데이터 자동 생성 방법에 대해 연구하였고 인공지능을 통해 자동 생성에 대한 가능성을 확인할 수 있었다.

추후 다양한 메타데이터 추출 요소에 대해 학습하고 자동 생성 시스템을 구축한다면 초개인화를 위한 맞춤형 추천 시스템을 통해 사용자에게 높은 만족도와 기업은 소비를 촉진 시킬 수 있는 서비스를 제공할 수 있을 것이다.

Acknowledgments

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No.2021R1I1A3057800) 과제 지원으로 연구되었으며, 관계부처에 감사드립니다.

References

[1] K. W. Song, and I. C. Moon, "Introduction to recent recommendation system research and future works," *Communications of the Korean Institute of Information Scientists and Engineers*, Vol. 39, No. 3, pp. 16-23, Mar. 2021.

[2] J. Y. Kim, and S. W. Lee, "The ontology based, the movie contents recommendation scheme, using relations of movie metadata," *Journal of Intelligence and Information Systems*, Vol. 19, No. 3, pp. 25-44, Sep. 2013.

[3] S. J. Bae, "Trend analysis of movie content curation and metadata standards research -focus on the art management perspective-," *Journal of Korea convergence society*, Vol. 11, No. 6, pp. 163-171, Jun. 2020.

[4] M. H. Yoon, H. G. Park, and I. Y. Moon, “A research of optimized metadata extraction and classification of in audio,” in *Proceeding of the 49th Conference on the Korea Institute of Information and Communication Engineering*, Yeo Soo: YS, KR, pp. 147-149, 2021.

[5] E. S. Noh, S. R. Yi, and S. M. Hong, “Binary classification of bolts with anti-loosening coating using transfer learning-based CNN,” *Journal of Korea Academia-Industrial cooperation Society*, Vol. 22, No. 2, pp. 651-658, Feb. 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas: LV, USA, pp. 770-778, 2016.

[7] E. J. Jeon, H. J. Kang, and C. J. Park, “Development of a Web-based class video editing program for automatic words and sentence segments deletion based on voice recognition,” *Journal of Korean Association of Computer Education*, Vol. 24, No. 3, pp. 45-56, May. 2021.

용 성 중 (Sung-Jung Yong)
2020년 8월 : 한국기술교육대학교 대학원 컴퓨터공학과 공학석사
2021년 8월 ~ 현재 : 한국기술교육대학교 대학원 컴퓨터공학과 박사과정
※관심분야 : AI, 빅데이터, 추천 시스템, 웹 등

박 효 경 (Hyo-Gyeong Park)
2021년 8월 :한국기술교육대학교 컴퓨터공학 학사
2021년 8월 ~ 현재 : 한국기술교육대학교 대학원 컴퓨터공학과 석사과정
※관심분야 : AI, 빅데이터, 추천 시스템 등

유 연 휘 (Yeon-Hwi You)
2016년 3월 ~ 현재 :한국기술교육대학교 컴퓨터공학 학사과정
※관심분야 : AI, 빅데이터, 추천 시스템 등

문 일 영 (Il-Young Moon)
2005년 2월: 한국항공대학교 항공통신정보공학과 공학박사
2005년 3월 ~ 현재: 한국기술교육대학교 컴퓨터공학과 교수
※관심분야 : AI, 무선인터넷 응용, 무선 인터넷, 모바일IP 등