Journal of KIIT. Vol. 23, No. 5, pp. 149-157, May 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.5.149

비주얼 리듬을 활용한 동영상 장르 분류 방법

정혜원*, 이해연**

Video Genre Classification Method using Visual Rhythm

Hye-Won Jung*, Hae-Yeoun Lee**

국립금오공과대학교 대학 연구과제비(2024~2025)로 지원되었음

요 약

OTT 서비스 및 동영상 플랫폼의 대중화로 인해 동영상 콘텐츠 소비가 증가하였고, 이에 따라 효과적인 콘텐츠 장르 분류의 중요성과 수요가 증가하였다. 이러한 사회적인 변화에 맞춰 최근 인공지능 기술을 통해 동영상 장르 분류를 수행하는 연구가 활발히 진행되고 있다. 본 논문에서는 딥러닝을 이용하여 빠르게 동영상 콘텐츠의 장르를 분류하는 방법을 제안한다. 제안하는 기법에서는 동영상 프레임의 시각적 특징을 반영한 비주얼 리듬을 생성하고, 이를 변형된 Densenet201 모델에 적용하여 방대한 크기의 동영상을 빠르게 장르 분류할 수 있다. 유튜브 등에서 수집한 다양한 동영상을 6개 장르로 분류하여 동영상 데이터셋을 구축한 후에 제안하는 방법에 대한 성능을 검증하였고, 98.5%의 장르 분류 정확도를 달성하였다.

Abstract

With the popularization of OTT services and video platforms, the consumption of video content has increased. Accordingly, the importance and demand for effective content genre classification have also grown. In line with these societal changes, recent research on video genre classification using artificial intelligence technology has been actively conducted. In this paper, we propose a method for rapidly classifying video content genres using deep learning. The proposed technique generates a visual rhythm that reflects the visual characteristics of video frames and applies it to a modified Densenet201 model for the fast genre classification of large-scale videos. After constructing a video dataset by categorizing various videos collected from platforms such as YouTube into six genres, we verified the performance of the proposed method and achieved a genre classification accuracy of 98.5%.

Keywords
video genre classification, visual rhythm, deep learning, modified Densenet201

* 국립금오공과대학교 컴퓨터공학부 학사과정
- ORCID: https://orcid.org/0009-0007-5563-2653
** 국립금오공과대학교 컴퓨터공학부 교수(교신저자)
- ORCID: http://orcid.org/0000-0002-6081-1492

· Received: Mar. 17, 2025, Revised: Apr. 07, 2025, Accepted: Apr. 10, 2025
· Corresponding Author: Hae-Yeoun Lee
Dept. of Computer Software Engineering, Kumoh National Institute of Technology, Korea
Tel.: +82-54-458-7548, Email: haeyeoun.lee@kumoh.ac.kr

Ⅰ. 서 론

현대 사회의 빠른 디지털화와 IT 기술의 발전으로 멀티미디어 콘텐츠의 소비와 수요가 증가하였다. 특히 OTT(Over-The-Top) 서비스와 유튜브 같은 동영상 플랫폼의 대중화는 동영상 콘텐츠의 생산과 소비를 크게 확대시켰으며, 이는 여가생활뿐만 아니라 광고, 교육, 정보 전달 등 다양한 분야에서 동영상 콘텐츠의 중요성 또한 증가시켰다[1]. 동영상 콘텐츠는 목적과 종류에 따라 다양한 장르로 구분될 수 있으며, 이러한 장르를 정확하고 효율적으로 분류하는 기술은 콘텐츠 관리, 개인화 추천 시스템, 맞춤형 광고, 그리고 사용자 경험 개선 등에서 핵심적인 역할을 한다.

기존 딥러닝 기반 동영상 장르 분류는 주로 프레임 차이나 자막 데이터를 활용한 다중 클래스 분류 문제를 다뤄왔다. 그러나 이는 동영상의 시각적 특징을 충분히 반영하지 못하는 것으로 보인다. 따라서 동영상의 프레임에서 시각적 특징을 추출하고 이를 기반으로 장르를 분류하는 것은 유용한 접근 방식이 될 수 있다.

본 논문에서는 동영상에 대하여 프레임 기반의 비주얼 리듬을 생성하고 딥러닝 모델을 통해 각 장르를 빠르게 분류하는 방법을 제안한다. 방대한 크기의 동영상 프레임들의 특징을 간소화하여 잘 표현하기 위하여 비쥬얼 리듬을 생성하였고, 딥러닝 모델의 분류 성능을 강화하기 위해 Densenet201 모델을 채택하여 변형과 파라미터 값 조정을 통해 최적화를 수행하였다.

제안하는 방법의 성능 평가를 위하여 6개 장르로 구분하여 유튜브 등에서 다양한 길이의 동영상을 수집한 후에 딥러닝 모델의 학습 효율성을 높이고 학습 조건의 일관성을 맞추기 위해, 수집한 동영상을 약 3분 길이로 나누어 동영상 데이터셋을 구축하였다. 이 데이터셋을 사용하여 제안한 방법의 장르 분류 학습과 성능 검증을 수행한 결과 98.5%의 분류 정확도를 달성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 동영상 장르 분류에 관한 국내외 연구를 요약하고, 3장에서는 프레임 기반 비주얼 리듬을 활용한 동영상 장르를 분류하는 방법을 제안한다. 4장에서는 제안한 방법의 성능 검증 결과를 기술하며, 5장에서는 논문을 결론짓는다.

Ⅱ. 관련 연구

국내외에서 수행되는 동영상 분류 연구는 동영상 종류 분류, 행동 인식 및 분석, 스포츠 유형 분류, 영화 장르 분류 등 다양한 분야에서 진행되고 있다. 특히, 최근 대부분의 연구에서는 딥러닝 기술을 적극적으로 활용하고 있다[2][3].

많은 연구에서 동영상의 핵심 프레임을 추출한 뒤, 프레임 특징을 추출하여 이미지 분류에 적합하게 변형하는 전처리를 거치며, 분류의 정확도를 높이기 위해 시각적 특징과 텍스트 특징을 결합하는 방법도 제안되었다[3]. 이 외에도 프레임의 특징에서 시공간적 특징을 학습하거나 이중 신경망 구조를 사용한 개선된 CNN(Convolutional neural network) 모델을 활용하는 방법 등이 제안되었다[4][5].

N. Pang et al.[3]은 짧은 영상의 분류를 위해 동영상의 핵심 프레임에서 TimeSformer 모델을 적용하여 시각적 특징을 추출하고, BERT(Bidirectional Encoder Representations from Transformers) 모델을 사용해 텍스트 특징을 추출한 후, NextVLAD를 이용해 각 특징을 클러스터링하고 가중치 부여하여 32개의 클래스로 분류하였다. 이 방법은 BOVText 데이터셋에서 87.6%의 정확도를 보였다.

F. Candela et al.[4]은 TV 프로그램의 장르 분류에 대해 두 가지 프레임을 제안하였다. 첫 번째는 SSIM(Structural Similarity Index Map)을 통해 주요 프레임을 감지하고, 이를 ResNet50으로 분류하는 방법이며, 두 번째는 동영상의 샷 경계 탐지한 후에 Densenet121 모델로 특징을 추출하고, Transformer 모델로 공간적 및 시간적 의존성을 학습하여 성능을 향상하는 방법이다. 이들은 방송 채널에서 추출한 동영상으로 5개와 15개 클래스로 분류했으며, 각 방법은 각각 95%와 87%의 정확도를 보였다.

T. Gao et al.[5]은 스포츠 동영상 분류를 위해 이중 신경망 구조를 사용한 개선된 CNN 모델을 제안했다. 퍼지 노이즈 감소 알고리즘으로 동영상 품질

을 향상시키고, 저해상도 데이터셋에서 87.25%, 고해상도 데이터셋에서 98.04%의 정확도를 달성했다.

E. Akarsu et al.[6]은 Anomaly Detection Dataset-UCF 동영상을 4개의 클래스로 분류하기 위해 전이 학습을 활용한 개선된 CNN 모델을 제안하였다. AlexNet, VGG19, ResNet18 모델을 사용해 동영상 프레임에서 특징을 추출한 뒤, LSTM(Long Short-Term Memory) 분류기를 활용하여 5-fold cross-validation 방식으로 분류하였다. 그 결과, AlexNet, VGG19, ResNet18은 각각 94.3%, 96%, 98%의 분류 정확도를 기록하였다.

III. 비주얼 리듬 기반 동영상 장르 분류 방법

본 절에서는 동영상에서 프레임 기반 비주얼 리듬을 생성하고, 이를 변형된 Densenet201 모델을 통해 동영상의 장르를 빠르게 분류하는 제안하는 방법을 설명한다. 이 방법은 3차원 동영상(2차원 프레임 + 시간축)을 2차원 비주얼 리듬 이미지로 변환함으로써 데이터 크기를 줄였기 때문에 빠른 분류 속도를 달성할 수 있다.

일반적으로 동영상은 제작자의 의도에 따라서 다양한 길이로 제작된다. 이와 같은 동영상에 대하여 1개의 고정 크기 비주얼 리듬을 생성하는 것은 시간적인 특성이 왜곡되고 해당 동영상의 장르 특징을 제대로 반영하지 못할 가능성이 있어서 장르 분류의 정확도가 낮아질 수 있다. 따라서, 제안하는 방법에서는 약 3분 정도의 동영상을 분류의 단위로 하여 장르를 분류한다. 만약에 3분 이상으로 재생시간이 긴 동영상에 대하여 제안하는 알고리즘을 적용하여 분류를 수행하려는 경우에는, 전체 동영상을 3분 길이 세그먼트로 구분한 후에 각 세그먼트에 대하여 제안하는 기법으로 분류를 수행한 후에 빈도수 등을 기반으로 최종 장르로 분류하는 형태로 변형하여 적용할 수 있다.

다음 절에서는 먼저 장르별 동영상을 수집한 후에 약 3분 길이 동영상 데이터셋의 구축에 대해서 설명하고, 그 후에 동영상에 대하여 프레임 기반 비주얼 리듬의 생성 방법과 이를 이용하여 변형된 Densenet201 모델을 통하여 장르 분류를 수행하는 방법을 설명한다.

3.1 장르별 동영상 데이터셋 구축

온라인에 공개된 동영상 데이터셋은 객체 인식 및 행동 분류에 초점을 둔 경우가 많으며, 이러한 데이터셋은 본 연구의 목적인 동영상 장르 분류에 직접 사용하기 어렵다[7]. 따라서, 장르별 동영상들의 특성을 정리해서 표 1과 같이 6개의 장르로 정의하고 동영상을 분류하고 수집하였다. 동영상들은 유튜브 등에서 다운로드하여 수집하였고, 다양한 길이를 갖고 있으며, 제안하는 방법의 학습이 편중되지 않도록 다양한 주제의 동영상을 선정하였다.

표 1. 동영상 클래스 정의와 특성
Table 1. Video class definition and characteristics

Video class	Characteristic	# of videos
Drama	* Close-ups on characters * Focusing on emotion	39
Entertainment	* Fast scene transitions * Dynamic Visual Effects	57
Movie	* Varied camera angles * Lighting and Color Use	40
News	* Static scene transitions * Text and graphic overlays	47
Sports	* Dynamic camera movement * Real-time data on screen	35
VLOG	* Natural camera movement * Casual scene transitions	37
Total	-	255

동영상 장르는 각 장르가 가지는 시각적 및 편집적 특성에 따라 분류될 수 있다.

드라마는 스토리 전개와 감정 표현을 강조하며, 주로 인물 클로즈업과 배경 전환이 특징적이다. 화면 전환은 서사 흐름에 맞춰 느리게 이루어지고, 대화 리듬과 캐릭터 관계를 나타내는 장면 전환도 중요한 역할을 한다. 영화는 장르와 주제 따라 다르지만, 일반적으로는 다양한 구도와 배경 전환을 통해 시각적인 몰입감을 제공하며, 조명과 색감의 활용이 두드러진다. 예능은 빠른 화면 전환과 역동적인 그래픽, 자막 등을 특징으로 하며, 리액션 컷과 웃음 포인트 편집이 자주 사용된다. 뉴스는 고정적이고, 안정적인 구도를 특징으로 하며, 배경은 주로 스튜디오 화면이나 현장 리포트 장면으로 구성된다.

또한, 정보 전달을 위해 헤드라인, 자막 등 그래픽 오버레이가 자주 사용된다. 스포츠는 경기의 역동성을 반영하기 위해 빠른 카메라 이동과 컷 전환이 빠른 카메라 이동과 컷 전환을 자주 사용된다. 또한 점수와 경기 상황과 같은 실시간 데이터가 화면에 표시되며, 특정 포인트에서는 반복 재생 및 슬로 모션이나 하이라이트 장면이 삽입되어 관람의 몰입도를 높인다. 브이로그는 일상적인 내용을 담은 영상으로, 자연스러운 카메라 움직임과 다양한 배경 변화가 특징이다.

다양한 길이의 수집된 장르별 동영상에 대해서 표 2와 같이 1,276개의 약 3분 길이 동영상으로 추출하여 동영상 데이터셋을 구축하였다. 수집된 동영상은 그 길이가 다양하므로 전체 길이에 따라서 추출할 동영상 개수를 다르게 설정하였는데, 3분 미만의 동영상은 원본을 그대로 사용했고, 3분 30초 미만의 동영상은 1개, 4개 미만의 동영상은 2개, 5분 20초 미만의 동영상은 3개, 6분 40초 미만의 동영상은 4개, 그 이상의 동영상은 5개로 추출했다.

표 2. 각 클래스에서 약 3분 길이 동영상 개수
Table 2. Total numbers of 3m video for each class

Video class	# of videos	# of 3m segment
Drama	39	194
Entertainment	57	291
Movie	40	192
News	47	241
Sports	35	165
VLOG	37	193
Total	255	1,276

3.2 프레임 기반 비주얼 리듬 이미지 생성

동영상 장르 분류를 위한 변형된 Densenet201 모델에 입력하기 위해 각 동영상에서 프레임을 이용하여 비주얼 리듬 이미지를 생성하는 과정을 그림 1에 나타내었다.

약 3분 길이 동영상에 대하여 동일한 시간 간격을 가지도록 linspace를 활용하여 총 256개의 프레임(256×256×3)을 선정하고, 각 프레임에 대해 RGB 각 채널별로 대각선 형태로 픽셀을 선정하여 비주얼 리듬 이미지를 추출하였다. 결과적으로 총 256개의 프레임에서 256×256×3 크기의 비주얼 리듬 이미지를 얻을 수 있다. 이와 같은 데이터 전처리 과정을 통하여 생성된 비주얼 리듬 이미지는 동영상 장르 분류를 위해 변형된 Densenet201 모델에 사용된다.

프레임 기반 비주얼 리듬 이미지는 대각선 패턴의 규칙성 및 밝기 변화를 통해 동영상 장르의 편집 스타일과 시각적 특징을 효과적으로 반영할 수 있다. 특히, 방대한 크기의 동영상을 1장의 이미지로 표현하여 빠르게 장르 분류를 수행할 수 있다.

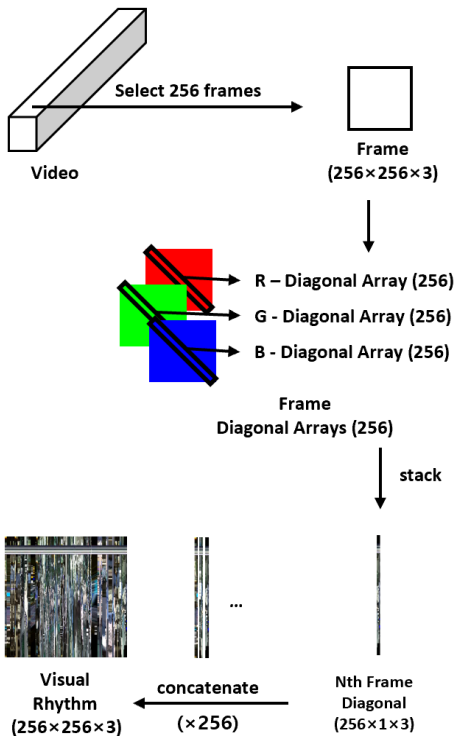

그림 1. 프레임 기반 비주얼 리듬 이미지 생성
Fig. 1. Visual rhythm image generation using frame

동영상 장르별로 비주얼 리듬 이미지를 생성한 결과는 그림 2에 나타내었다.

드라마의 경우 인물 클로즈업과 배경 변화로 인해 비주얼 리듬 이미지에서 대각선 밝기의 변화가 뚜렷하게 나타나며, 변화 간격이 비교적 넓게 보인다. 영화의 경우는 장르와 주제에 따라 다양한 패턴을 보인다. 빠른 장면 전환이 많은 액션 영화는 불규칙적인 대각선 패턴을 형성하며, 서사의 흐름이 느린 영화는 드라마와 유사한 규칙성을 나타낸다.

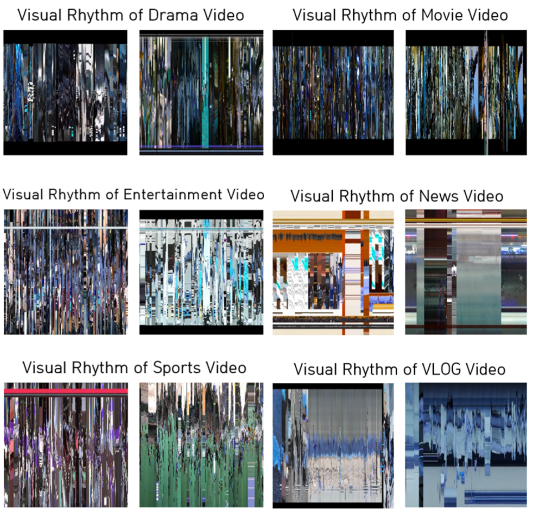

그림 2. 장르별 비주얼 리듬 이미지의 차이
Fig. 2. Difference of visual rhythm image by genre

예능은 화려한 그래픽 요소와 높은 화면 변화율로 인해 역동적인 대각선 패턴과 강한 밝기 변화를 보인다. 뉴스는 고정된 구도와 안정적인 화면 구성을 특징으로 하여 단조롭고 반복적인 형태의 패턴을 보인다. 스포츠의 경우 빠른 카메라 이동과 잦은 장면 전환으로 인해 대각선 패턴이 빠르게 변화하며, 경기 상황과 점수 표시와 같은 실시간 정보로 인한 불규칙적인 흐름도 보인다. 브이로그는 비교적 규칙적인 패턴을 보이지만, 배경 변화로 인해 패턴 진행의 변화를 보이기도 한다.

3.3 변형된 Densenet201 모델

딥러닝은 방대한 데이터에서 의미 있는 특징을 스스로 학습하는 기술로, 여러 개의 계층과 은닉층을 쌓아 정교하고 깊은 모델을 만들어서 정확도를 향상한다. 특히, 하드웨어와 GPU의 성능 발전에 힘입어 딥러닝이 발전되면서 컴퓨터 비전 분야에서 각광받고 있다. 딥러닝에서 이미지 분류 모델과 알고리즘이 발전하면서 동영상 분류 연구로 확장이 진행되고 있다. 동영상의 시간적 특성을 그대로 입력으로 사용하는 경우도 있지만, 적절한 전처리를 통해 동영상의 시각적 특징과 텍스트 특징을 이용하여 분류하는 알고리즘도 제안되고 있다.

Densenet201은 심층 CNN의 한 종류로서, 총 201개의 계층으로 구성된 신경망 모델이다[8][9]. 이 모델은 Dense Block 구조와 Transition Layer를 기반으로 설계되었으며, 입력 데이터를 처리하는 1개의 입력 계층, 특징 추출 및 학습을 수행하는 199개의 계층, 그리고 결과를 판단하는 1개 분류 계층으로 이루어져 있다.

Densenet201 모델의 핵심은 계층 간 밀집 연결 구조로서 각 계층의 출력을 이후 모든 계층의 입력으로 사용하는 형태로 연결되어서 다양한 특징들이 지속적으로 전달되어 효과적인 특징 추출이 가능하다. 이러한 설계는 기존 CNN에서 깊이가 길어질수록 발생하던 저수준 특징의 소실 문제를 완화하고, 밀집 연결 구조를 통해 기울기가 원활하게 전달되어, 오차 역전파 과정에서 발생할 수 있는 기울기 소실 문제를 줄여준다. 또한, Densenet은 여러 계층으로 구성된 Dense Block을 통해 연결 연산을 수행한다. Dense Block은 계층 간 밀집 연결을 통해 그래디언트 소실 문제를 완화하고, 모델이 풍부한 특징을 효과적으로 학습할 수 있도록 설계되었다. 또한, Growth rate를 조절하여 각 계층이 학습에 기여하는 정도를 결정한다. Dense Block 사이에는 Transition 계층이 위치하며, Pooling 연산을 통해 특징맵의 크기를 줄이고 연산량을 최적화한다. 특히, Bottleneck 계층은 입력값 채널을 줄여 연산량을 경량화시킨다. 마지막 계층에서는 전역 평균 풀링(Global average pooling)을 사용하여 추출된 특징맵을 1차원 벡터로 변환하고, 이를 전연결(Fully connected) 계층과 연결한다. 이후, 소프트맥스 계층을 통해 최종적으로 주어진 입력을 분류한다[8][9].

기존 Densenet201 모델은 ImageNet 데이터셋에 최적화된 딥러닝 모델로서 본 연구에서는 동영상 장르 분류를 위해서 Densenet201 모델을 변형하여 활용하였다. 이 변형된 Densenet201 모델은 밀집 연결(Dense connection) 구조를 가지며, 모델의 구조를 그림 3에 간략하게 나타내었다.

그림에서 확인할 수 있듯이 변형된 Densenet201에서는 기본 Densenet201 모델의 제일 마지막에 위치한 1000D 전연결 계층을 제거하였고 Custom Classification Layer를 추가하였다.

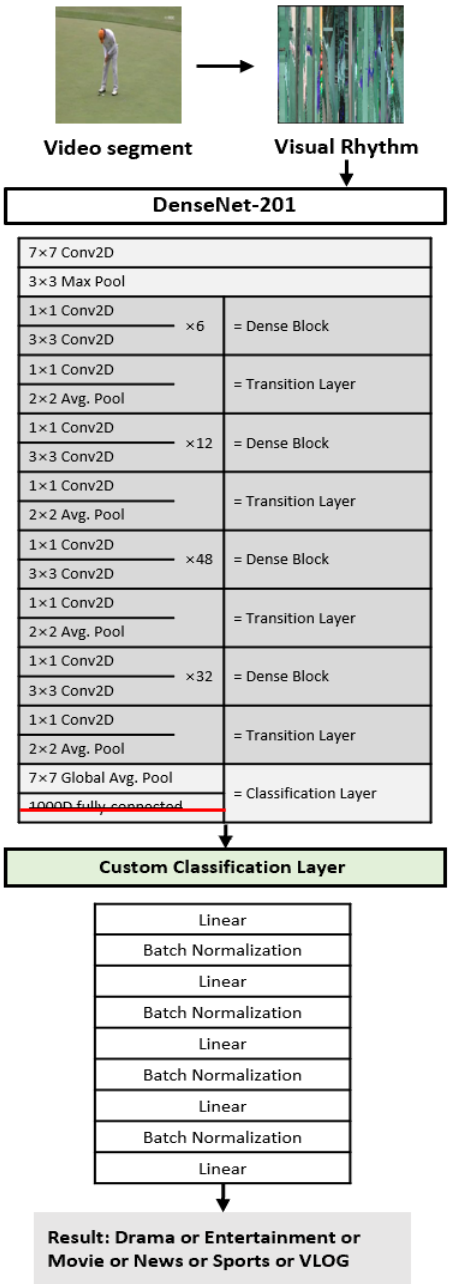

그림 3. 동영상 장르 분류를 위한 변형된 Densenet201 모델

Fig. 3. Modified Densenet201 model for video genre classification

추가된 Custom Classification Layer에는 4개의 선형화 계층과 배치 정규화를 수행하는 계층이 포함된다. 또한, 1개의 선형화 계층을 더 추가하여 모델이 장르별 동영상의 다중 분류 작업을 효과적으로 수행할 수 있도록 최적화하였다. 그리고, 다중 분류 작업에 사용되는 CrossEntropyLoss 손실 함수는 내부적으로 소프트맥스 함수를 포함하고 있으므로, 별도의 소프트맥스 변환 계층을 추가하지 않았다[7]. 참고로 소프트맥스 함수는 함수에 들어오는 값을 0과 1사이의 확률값으로 변환을 수행하며, 변환된 결과에 대한 합계가 1이 되도록 처리를 수행한다.

이처럼 변형된 Densenet201 모델은 3.2절에서 설명한 전처리 과정을 통해 생성된 비주얼 리듬 이미지를 입력으로 받아서 동영상 장르에 대한 분류를 수행한다.

Ⅳ. 실험 결과 및 분석

본 연구에서 제안하는 방법은 Intel i7-7700 CPU, nVidia RTX 2080 GPU, 64GB RAM, 그리고 Window 10 Pro 운영체제에서 개발되었고, 정확도 분석과 실험을 수행했다. 변형된 Densenet201 모델은 pyTorch를 활용하여 구현하였으며, 데이터셋 학습은 torch.utils.data API를 적용하여 지연 로딩(Lazy Loading) 방식을 사용했다. 이를 통해 제한된 메모리로 효율적인 학습이 가능하며, 필요한 데이터만을 GPU 메모리에 읽어들이기에 모델의 깊이와 입력 데이터 차원의 크기를 높일 수 있다.

4.1 실험 데이터셋과 변형된 Densenet 학습

동영상 프레임 기반 비주얼 리듬을 활용한 변형된 Densenet201 모델의 성능을 검증하기 위하여, 3.1절에서 설명한 것과 같이 구축한 약 3분 길이 동영상 데이터셋을 사용하였으며 표 3과 같이 실험을 위해 동영상 데이터셋에서 클래스별로 학습 60%, 검증 20%, 평가 20% 비율로 나누어 무작위로 순서를 섞어서 사용했다.

변형된 Densenet201 모델의 학습률은 10e-3, 16 배치 크기를 가지도록 모델 학습 파라미터를 설정하였고, 손실 함수는 소프트맥스 함수와 크로스 엔트로피 손실을 결합한 CrossEntropyLoss 함수, 옵티마이저는 ADAM(Adaptive Moment Estimation)을 사용하여 모델 학습을 진행하였다[10].

표 3. 각 클래스의 학습, 검증, 평가 동영상 개수
Table 3. Numbers of training, validation and testing video for each class

Class	Training	Validation	Testing	Total
Drama	116	38	40	194
Entertainment	174	58	59	291
Movie	115	38	39	192
News	144	48	49	241
Sports	99	33	33	165
VLOG	115	38	40	193
Total	763	253	260	1,276

학습은 최대 300 epoch까지 진행하지만 25 epoch 동안 검증 데이터의 손실이 감소하지 않으면 최적으로 판단하고 훈련을 마치도록 설정하였다.

4.2 동영상 장르 분류 실험 결과

4.1절의 실험 데이터셋을 이용하여 프레임 기반 비주얼 리듬을 생성한 후에 변형된 Densenet201 모델에 적용하여 학습과 검증을 수행하였고, 그 과정에서 손실 및 정확도 추세를 epoch 단위 그래프로 정리하여 그림 4에 나타냈다. 그림에서 X축은 epoch이고, Y축은 각각 손실값 및 정확도를 의미한다.

학습 과정에서 epoch가 진행됨에 따라 성능이 향상되는 것을 확인할 수 있다. 변형된 Densenet201 모델은 학습 진행 중 100 epoch에서 조기 종료되었으며, 61 epoch에서 검증 데이터셋 기준 98.02% 정확도를 보이며 최고 성능을 달성했다. 학습 과정에서 0 epoch에서 20 epoch까지는 검증 정확도가 급격히 상승하며 큰 그래프 변화를 보이지만, 20 epoch 이후에는 약 95% 전후에서 비교적 안정적인 정확도를 유지하는 경향을 보였다.

동영상 장르 분류를 위한 변형된 Densenet201 모델이 학습 중 과적합되지 않고 안정적인 학습이 이루어졌음을 확인하기 위해 평가 데이터셋을 이용하여 성능 평가를 수행하였다. 학습 과정에서 가장 낮은 검증 손실값을 기록한 최적 모델인 75 epoch 모델을 사용하였으며, 평가 데이터 260개를 대상으로 테스트한 결과 4개의 분류 오류가 발생하여 98.5%의 정확도를 달성하였다.

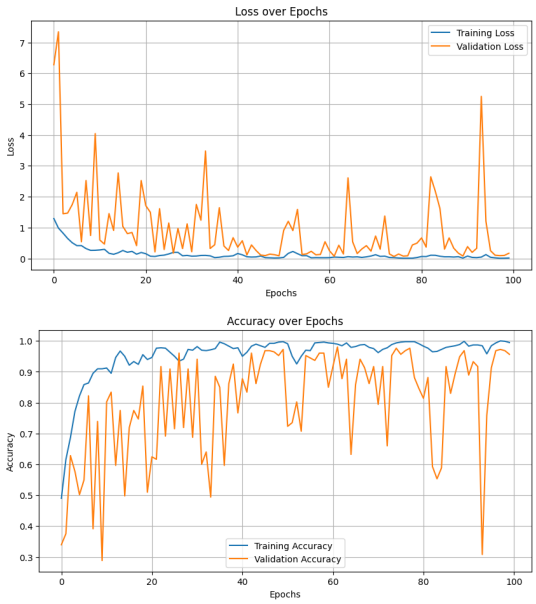

그림 4. 학습과 검증에서의 손실 및 정확도 추세
Fig. 4. Loss and accuracy trend of training and validation

그림 5는 동영상 장르 분류 결과를 혼동 행렬로 나타낸 것이며, 총 4개의 분류 오류가 발생하였다. 평가 데이터셋에서 드라마 2개가 영화로, 예능 2개가 각각 스포츠와 브이로그로 잘못 분류된 것을 확인할 수 있다.

일반적으로 드라마는 인물 클로즈업과 감정 표현을 강조하며, 시간에 흐름에 따라 비교적 느린 화면 전환을 보이는 반면, 영화는 다양한 구도와 배경 전환을 활용하여 시각적 몰입감을 제공한다.

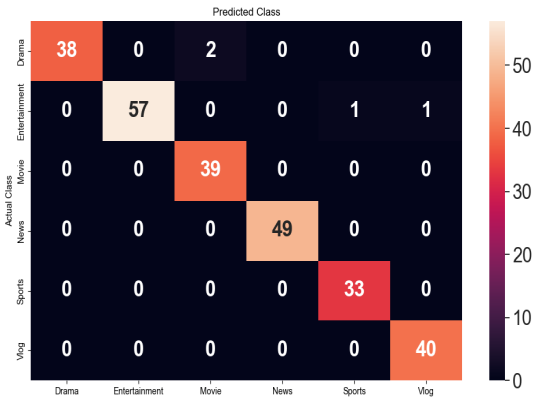

그림 5. 평가 데이터에 대한 혼동 행렬
Fig. 5. Confusion matrix for testing dat

그러나, 영화로 분류된 드라마는 인물이 처한 상황을 극대화하기 위해 영화에서 주로 사용되는 다양한 구도와 배경 전환 기법을 활용하고 있어, 영화의 비주얼 리듬 패턴과 유사하게 나타난 것으로 판단된다.

또한, 예능의 경우 콘텐츠의 특성에 따라 스포츠 또는 브이로그로 분류하는 경우가 발생하였다. 스포츠로 분류된 예능은 헬스장에서 운동하는 상황을 다루고 있어, 스포츠 장르의 역동적인 장면과 유사한 패턴을 포함하고 있는 것으로 판단된다. 브이로그로 분류된 예능의 경우, 특정 인물의 몸짓에 집중하는 콘텐츠로, 다른 예능들에 비해 상대적으로 장면 전환과 시각적 변화가 적고 자연스러운 카메라 움직임을 포함하고 있어 브이로그와 유사한 비주얼 리듬 패턴을 보였을 것으로 판단된다.

이처럼 본 실험에서 발생한 분류 오류는 일부 장르 간 비주얼 리듬 특징이 겹치거나 특정 콘텐츠가 다른 장르의 특성과 유사하게 나타나면서 발생한 것으로 분석된다.

V. 결 론

멀티미디어 플랫폼의 대중화로 인해 다양한 장르의 동영상이 유통되면서, 동영상의 장르를 분류하는 기술에 대한 수요가 증가하고 있다. 동영상 분류를 위해 딥러닝을 활용한 연구가 활발하게 진행되고 있으며, 동영상의 시각적 특징뿐만 아니라 텍스트 특징 및 시공간적 특징을 이용한 분류 방법들이 제안되고 있다.

본 논문에서는 동영상의 프레임 기반 비주얼 리듬을 생성하고, 변형된 Densenet201 모델을 활용하여 시각적 특징을 추출한 뒤 이를 통해 동영상 장르를 분류하는 방법을 제안하였다. 유튜브 등에서 동영상을 수집하였고, 길이가 다양한 동영상을 동일한 조건에서 활용하기 위해 3분 단위로 분할하여 동영상 데이터셋을 구축하였다. 이 동영상에서 프레임 기반 비주얼 리듬 이미지를 생성하고, 이를 변형한 Densenet201 모델에 활용하여 학습과 분류를 진행하였다. 성능 검증을 위한 실험에서 제안하는 방법은 98.5% 분류 정확도를 달성하였으며, 동영상 장르 분류를 위한 효과적인 접근 방식임을 보여주었다. 제안하는 방법은 방대한 동영상 데이터를 빠르게 장르별로 분류하여 멀티미디어 플랫폼의 증가하는 수요에 부응하고, 사용자 편의성을 크게 향상하는데 기여할 수 있는 것으로 기대된다.

제안한 방법에서는 프레임을 기반으로 각 장르의 장면 전환 및 시각적 요소를 중심으로 특징을 추출하여 동영상 장르를 분류하였다. 전체적으로 높은 분류 정확도를 보였으나, 몇 가지 분류 오류가 확인되었다. 드라마와 영화의 경우, 스토리에 따라 화면 구성과 장면 전환 패턴이 유사하여 시각적 특징을 공유하는 경향이 있어 잘못 분류되는 경우가 발생하였다. 또한, 예능 동영상의 경우, 운동을 다루는 콘텐츠는 스포츠 동영상으로, 자연스러운 일상 형식의 콘텐츠는 브이로그로 잘못 분류되는 오류가 관찰되었다. 이러한 한계를 극복하기 위해서는 새로운 특징 분석 및 추출 방법을 도입하거나, 해당 장르의 시각적 특징을 보다 명확히 구분하기 위한 개선된 모델과 방법을 제안하는 후속 연구가 필요할 것으로 보인다. 또한, 길이가 무작위인 동영상에도 제안하는 방법을 적용하기 위하여 3분 길이 세그먼트로 장르를 분류한 후에 빈도수 등을 활용하여 최종 장르를 판정하는 형태로 연구를 확장할 계획이다.

References

[1] Y. Kim, "Who Uses OTT Services and How Much Time They Spend There? Focusing on Socio-Demographic and Personality Characteristics", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 23, No. 5, pp. 29-34, Oct. 2023. http://dx.doi.org/10.7236/JIIBC.2023.23.5.29.

[2] E. Lee and S. Han, "Comparison of Loss Function for Multi-Class Classification of Collision Events in Imbalanced Black-Box Video Data", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 24, No. 1, pp. 49-54, Feb. 2024. http://dx.doi.org/10.7236/JIIBC.2024.24.1.49.

[3] N. Pang, S. Guo, M. Yan, and C. A. Chan, "A Short Video Classification Framework Based on Cross-Modal Fusion", Sensors, Vol. 23, No. 20, pp. 8425, Oct. 2023. https://doi.org/10.3390/s23208425.

[4] F. Candela, A. Giordano, C. F. Zagaria, and F. C. Morabiot, "Effectiveness of deep learning techniques in TV programs classification: A comparative analysis", Integrated Computer-Aided Engineering, Vol. 31, No. 4, pp. 439-453, Nov. 2024. https://doi.org/10.3233/ICA-240740.

[5] T. Gao, M. Zhang, Y. Zhu, Y. Zhang, X. Pang, J. Ying, and W. Liu, "Sports Video Classification Method Based on Improved Deep Learning", Applied Sciences, Vol. 14, No. 2, pp. 948, Jan. 2024. https://doi.org/10.3390/app14020948.

[6] E. Akarsu and T. Karacali, "Comparison of Video Classification Results with Machine Learning", Traitement du Signal, Vol. 41, No. 2, pp. 847-855, Apr. 2024. https://doi.org/10.18280/ts.410225.

[7] H. Choi, J. Kim, H. Hwang, and H.-Y. Lee, "Edited and Non-edited Video Classification Method through Histogram-based Shot Change Detection", Proc. of the Korea Digital Forensic Society, Seoul, Korea, 2023.

[8] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks", Proc. of the IEEE conference on computer vision and pattern recognition, Honolulu Hawaii, pp. 4700-4708, Jul. 2017. https://doi.org/10.48550/arXiv.1608.06993.

[9] Y.-E. Seo, S.-G. Ahn, and H.-Y. Lee, "MFCC-based Audio Quality Classification Method for Forensics", Journal of KIIT, Vol. 22, No. 1, pp. 131-138, Jan. 2024. http://dx.doi.org/10.14801/jkiit.2024.22.1.131.

[10] T. Dozat, "Incorporating Nesterov Momentum into Adam", Proc. of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, pp. 1-4, May 2016.

저자소개

정 혜 원 (Hye-Won June)

2023년 3월 ~ 현재 : 국립금오공과대학교 컴퓨터소프트웨어공학과 학사과정

관심분야 : Computer Vision, Deep Learning

이 해 연 (Hae-Yeoun Lee)

1997년 : 성균관대학교 정보공학과 (공학사)

1999년 : KAIST 전산학과 (공학석사)

2006년 : KAIST 전자전산학과 (공학박사)

2008년 ~ 현재 : 국립금오공과대학교 컴퓨터소프트웨어공학과 교수

관심분야 : Digital Forensics, Image Processing, Computer Vision, Internet of Things