의미 기반 영화 장면 검색 시스템

무비탐정 16기 박지혜, 16기 박주연, 16기 남경민

1. 연구 배경 및 목적

배경

영화는 영상, 대사, 음악, 카메라 연출이 동시에 작동하는 멀티모달 콘텐츠이며, 사용자가 장면을 찾을 때는 단순 키워드보다 상황이나 감정을 기준으로 탐색하는 경우가 많다. 예를 들어 사용자는 "주인공이 죄책감을 드러내는 장면"이나 "가족이 한 방에서 어색하게 침묵하는 장면"처럼 서사적 의미를 질의한다. 그러나 기존의 보편적인 검색 방식은 제목이나 장르 같은 정적 메타데이터, 자막 텍스트의 키워드 매칭, 사람이 수동으로 붙인 태그에 크게 의존한다.

이 방식은 대사와 같이 명시적으로 등장하지 않는 추상적 개념에 대한 질의에 대한 검색을 수행하기 어렵다는 한계점을 가진다. 즉, 의미 기반 검색을 수행하기 위해서는 인물의 표정, 자세와 같은 시각 단서(Visual), 인물의 대사 및 호칭 등의 언어 단서(Language), 감정적 톤이나 효과음의 청각 단서(Audio)를 풍부하게 활용하여 대상 데이터셋과의 비교할 수 있어야 한다.

목적

본 연구의 최종 목적은 사용자 한국어 질의를 입력으로 받아, 영화 데이터셋 내부에서 의미적으로 관련성이 높은 장면을 자동으로 찾아 영상 자료와 함께 정확한 시간 구간을 제공하는 검색 아카이브를 구축하는 것이다. 현실적으로 영상 데이터는 방대하기에 수작업으로 모든 장면에 태그를 달거나 설명을 작성하는 것은 비용과 시간적 측면에서 불가능에 가깝다, 따라서 본 프로젝트는 다음 문제의식에서 출발한다.

먼저 대규모 영상에서 장면을 자동으로 분할(Scene Segmentation)하고 각 장면의 프레임을 기반으로 적용한 등장인물 라벨링과 대사에 대한 STT 데이터를 결합하여 사용자의 한국어 질의에 대한 의미적으로 가장 유사한 Top-K 장면과 시간 구간을 반환하는 "의미 기반 영화 장면 검색 아카이브 시스템"이 필요하다고 판단했다.

본 연구는 영상의 시간 축을 임의로 자르지 않고, PySceneDetect 기반 샷 경계를 기준으로 데이터 단위를 통일한다. 이 기준 위에서 장면별 대표 프레임을 추출하고 프레임이 속한 장면 시간 구간을 기준으로 오디오 구간을 매핑하여 Captioning과 STT 데이터를 각각 추출하여 결합한다. 이 과정은 프레임 분할과 오디오 분할의 불일치로 생기는 정보 손실을 줄이고, 검색 결과를 정확한 시간 구간으로 환원하는 데 중요한 기반이 된다.

가설

	추가된 의미 정보	비교 대상	기대 효과
가설1	인물 라벨링	대조군 → 실험군 1	인물 기반 장면 구분 가능
가설2	STT 결합	실험군 1 → 실험군 2	장면 맥락 이해 범위 확장

본 연구는 대조군과 두 가지 실험군의 수행 결과를 비교하여, 장면 의미 표현을 풍부하게 만들수록 의미 기반 장면 검색 성능이 향상된다는 중심 가설을 검증한다.
인물 라벨링을 통해 시각적으로 유사하지만 등장인물이 다른 장면을 구분할 수 있어, 인물 관계를 포함한 질의에 대해 보다 정확한 장면 매칭이 가능할 것으로 예상한다(H1).
또한 캡션에 STT를 결합함으로써 대사 기반 맥락 단서가 추가되어, 시각 정보만으로는 포착하기 어려운 장면의 의미와 상황적 흐름을 더 폭넓게 반영할 수 있을 것으로 기대한다(H2).

2. 연구 방법

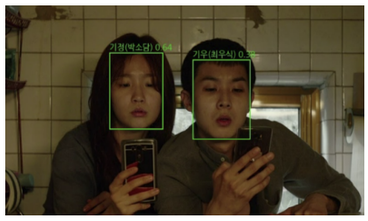
1. 데이터

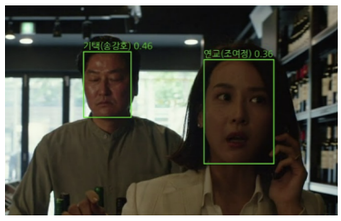
1.1 장면 단위 분할 (Scene Segmentation)

PySceneDetect를 활용하여 장면 전환 지점을 자동 감지하고, 영화를 998개의 scene clip으로 분할하였다. 분할된 각 장면은 이후 캡셔닝, 임베딩, 검색의 기본 단위로 사용된다.

1.2 인물 정보 수집 및 얼굴 기반 인식 (Character Preprocessing)

영화진흥위원회(KOFIC) 공식 사이트에서 주연 배우의 이름과 이미지를 크롤링하였다. 수집된 얼굴 이미지는 InsightFace를 통해 얼굴 임베딩으로 변환되었으며, 이를 기반으로 각 장면에서 인물의 얼굴을 바운딩 박스로 검출하였다. 해당 정보는 장면별 등장 인물 메타데이터로 활용된다.





1.3 음성 기반 정보 전처리 (Audio Preprocessing)

장면의 의미를 보완하기 위해 음성 정보 또한 분리하여 처리하였다. Whisper-large-v3 모델을 사용해 인물 대사를 텍스트로 변환하였으며, panns_inference(Pretrained Audio Neural Networks)를 활용해 배경음 및 환경음을 분석함으로써 장면의 분위기와 감정적 특성을 보조적으로 반영하였다.

2. 멀티모달 캡셔닝 기반 장면 의미 표현

2.1 이미지 캡셔닝 (Image-level Captioning)

바운딩 박스가 포함된 이미지를 Gemini API를 활용한 이미지 캡셔닝을 추출하였다. 이미지 캡셔닝은 일반적으로 비전 인코더를 통해 시각 정보를 임베딩 공간으로 변환한 뒤, 언어 디코더를 통해 자연어 설명을 생성하는 구조를 따른다. 이를 통해 장면 내 인물, 객체, 공간과 같은 정적인 시각 요소를 텍스트로 표현하였다.

2.2 비디오 캡셔닝 (Video-level Captioning)

바운딩 박스가 초함된 짧은 영상의 캡셔닝을 추출하였다. 비디오 캡셔닝은 시간 축을 고려한 시각 인코딩을 기반으로 인물의 행동, 장면 전개, 상황 변화와 같은 동적 정보를 텍스트로 생성한다. 이러한 방식은 CLIP 계열 모델에서 사용되는 비전–언어 공동 임베딩 개념과 유사하게, 시각 정보와 언어 표현을 동일한 의미 공간에서 연결하는 원리를 따른다.

3. 의미 기반 장면 검색 및 재생 시스템

3.1 문장 임베딩 모델

한국어 자연어 질의와 장면 텍스트를 동일한 의미 공간으로 매핑하기 위해 다국어 문장 임베딩 모델(multilingual-e5-large)을 사용하였다. 해당 모델은 검색 태스크에 특화된 query–passage 구조를 따른다.

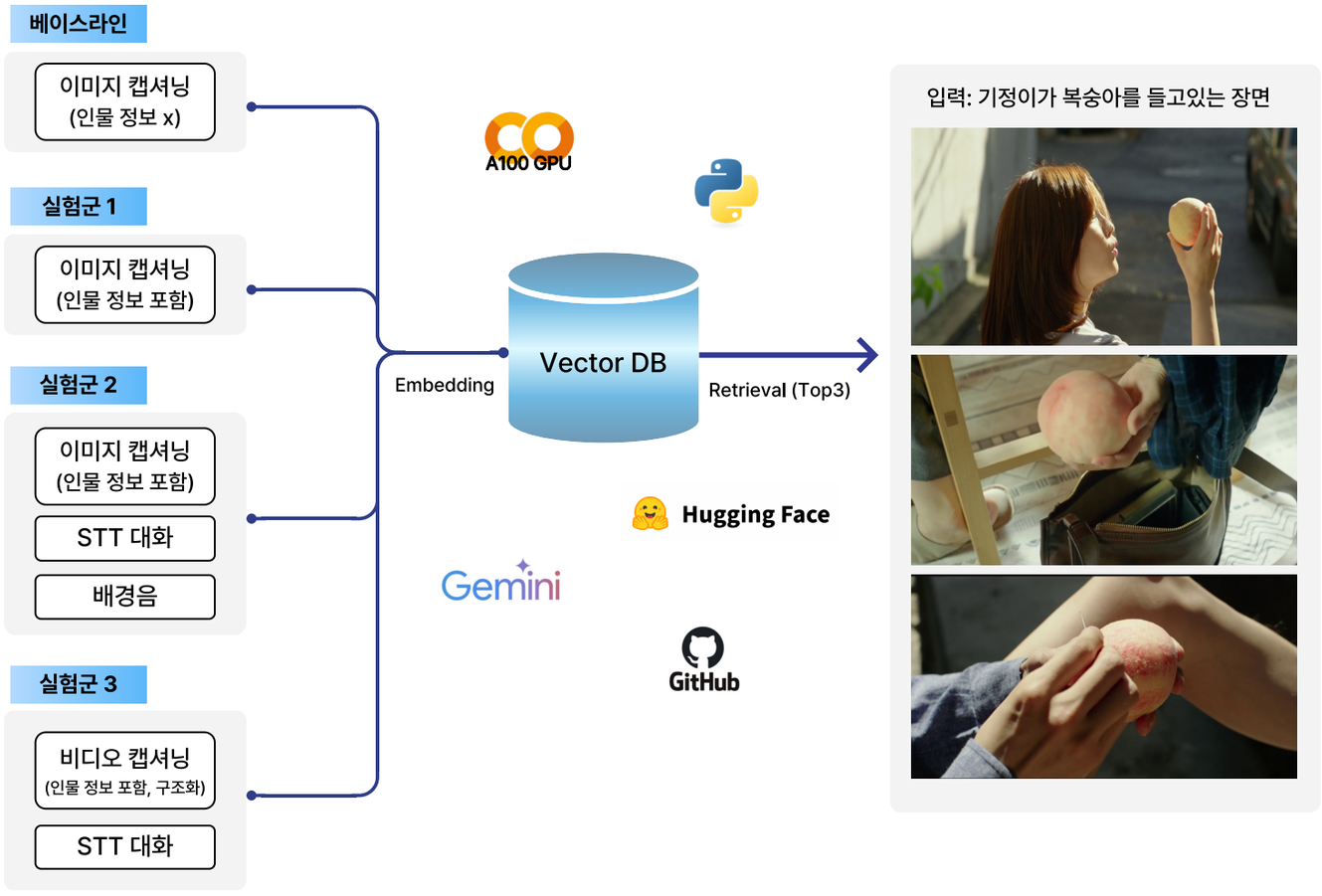
3.2 가중치 기반 의미 결합

시각 정보와 음성 정보에서 생성된 임베딩은 가중합 방식으로 결합된다. 이를 통해 장면 의미 표현 시 각 정보 축의 중요도를 조절할 수 있도록 설계하였다.

3.3 벡터 기반 검색 구조

모든 장면 임베딩은 FAISS (Facebook AI Similarity Search)기반 벡터 인덱스로 구성되며, 사용자 질의는 유사도 기반 Top-K 검색을 통해 의미적으로 가장 관련성 높은 장면을 검색한다.

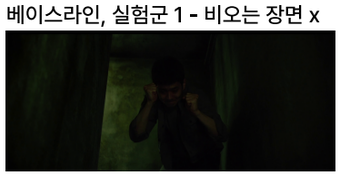
3. 연구 과정



4. 연구 결과

"비오는 날에
계단 내려가는 장면"

베이스라인, 실험군 1 - 비오는 장면 x



실험군 2, 3

"연교랑 기택이 차 안에서
대화하는 장면"

베이스라인 - 인물 필터링 불가능

실험군 1 - 직접적인 대화 x

실험군 2, 3

기존의 키워드나 시간대 중심 검색 방식과 달리, 본 시스템은 장면의 시각적 설명과 대사 정보를 임베딩 기반으로 통합하여 비교함으로써 장면이 담고 있는 상황과 맥락을 보다 정확하게 반영할 수 있음을 확인하였다. 또한 검색 결과를 사전에 분할된 scene clip으로 즉시 재생할 수 있도록 설계함으로써 사용자가 검색 결과를 직관적으로 검증할 수 있는 시스템을 구현하였다.

그러나 본 연구에는 몇 가지 **한계점**이 존재한다.
1. 캡셔닝 및 음성 인식 결과의 품질이 장면 의미 표현의 정확도에 직접적인 영향을 미친다는 점에서, 자동 생성된 텍스트의 오류가 검색 성능에 영향을 줄 수 있다.
2. 실험 대상이 단일 영화로 제한되어 있어 다양한 장르나 연출 스타일을 갖는 영화에 대한 일반화 성능은 추가적인 검증이 필요하다. 또한 인물 인식 과정에서 얼굴 가림이나 조명 변화가 큰 장면에서는 인물 식별 정확도가 낮아지는 한계가 존재한다.

향후 연구에서는 멀티모달 임베딩 모델을 활용한 장면 표현 고도화, 장르별 가중치 자동 조절, 그리고 사용자 질의 유형에 따른 검색 전략 개선을 통해 검색 성능을 향상시킬 수 있을 것으로 기대된다. 더 나아가 본 시스템은 대규모 영상 콘텐츠를 보유한 OTT 플랫폼, 예를 들어 Netflix와 같은 서비스에서 장면 검색, 하이라이트 추천, 콘텐츠 탐색 보조 기능 등으로 확장 가능성이 있으며, 영상 아카이빙 및 콘텐츠 분석 분야에서 실질적인 활용 가능성을 가진다.