특집논문 (Special Paper)
방송공학회논문지 제29권 제3호, 2024년 5월 (JBE Vol.29, No.3, May 2024)
https://doi.org/10.5909/JBE.2024.29.3.263
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

생성형 이미지를 이용한 멀티모달 장면 검색 시스템

이 주 희[a], 양 효 걸[b], 강 제 원[a], 한 치 영[b], 염 규 현[b]‡

Multi-modal Video Retrieval System using Generated Image Cue

Juhee Lee[a], Hyogeol Yang[b], Jewon Kang[a], Chiyeong Han[b], and Kyuhyun Yum[b]‡

요 약

최근 폭발적으로 증가하는 영상 콘텐츠로 인해 대량의 영상 데이터에 효과적으로 접근하기 위한 지능형 검색 기술이 필수적이다. 하지만 기존의 키워드 기반 영상 검색 기술은 한계가 존재한다. 사전에 추출된 메타 데이터와 키워드에 의존하기 때문에 복잡한 상황이나 추상적인 개념을 표현하기 어려웠고, 사용자 의도를 정확히 반영하지 못하는 경우가 많았다. 이에 본 연구에서는 생성형 AI 기술과 멀티모달 사전학습 모델을 융합한 새로운 영상 검색 시스템을 제안한다. 생성형 이미지 모델을 통해 텍스트로 표현하기 어려운 복잡한 상황이나 추상적 개념을 구체화한 가이드 이미지를 생성할 수 있다. 또한 대규모 데이터로 사전 학습된 멀티모달 모델을 활용하여 텍스트, 이미지, 비디오 데이터를 동일 의미 공간에 정렬시켜 유사도를 측정한다. 제안 기술에서는 사용자 텍스트 입력과 생성 이미지 가이드를 조합하여 지능형 시멘틱 검색을 수행한다. 이를 통해 기존 방식이 가진 한계를 극복하고 사용자의 의도에 부합하는 검색 결과를 제공할 수 있다. 본 연구는 생성형 AI 기술이 영상 검색과 결합했을 때, 새로운 시너지가 발생할 수 있음을 보여주고 있으며, 새로운 방식의 멀티모달 영상 검색 기술은 방송사 아카이브 관리와 콘텐츠 제작 분야에서 큰 효율성과 생산성 향상을 가져올 것으로 기대된다.

Abstract

In this paper. we introduce a novel video retrieval system utilizing generative image models and multi-modal pre-trained models. This method uses large-scale datasets to align video and text data into the same semantic space using pre-trained models, enabling effective retrieval of videos with semantic similarity to text queries. Particularly, it incorporates an efficient semantic search method that can reflect abstract concepts difficult to express in text, guided by images. The proposed method is expected to play a crucial role in enhancing the efficiency of video scene retrieval and management dealing with vast amounts of video data in the content production system and market.

Keyword : Text to video retrieval, Multi-modal search, Generative model

a) 이화여자대학교 공과대학 전자전기공학과(Department of Electronic and Electrical Engineering, Ewha Womans University)
b) 문화방송 AI전략자회사TF(MBC AI Strategy subsidiary TF)
‡ Corresponding Author : 염규현(Yum, Kyuhyun)
E-mail: email@mbc.co.kr
Tel: +82-2-789-2549
ORCID: https://orcid.org/0009-0004-6243-1284
※ 이 연구는 과학기술정보통신부 재원으로 한국전파진흥협회(RAPA)의 지원을 받아 수행된 연구임.
This work was supported by Korea Radio Promotion Association(RAPA) grant funded by The Ministry of Science and ICT
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음(IITP-2024-2020-0-01460)
· Manuscript April 4, 2024; Revised May 1, 2024; Accepted May 2, 2024.

Ⅰ. 서 론

현대 미디어 사업은 소셜 미디어 등의 다양한 플랫폼 등을 통해 막대한 양의 이미지와 비디오 콘텐츠를 생성하고 유통한다. 콘텐츠 시장이 방대해짐에 따라 대량의 영상 자료에 신속하고 효과적으로 접근할 수 있는 지능형 검색 기술의 중요성이 증대되었다. 특히 멀티모달 학습 모델의 발전으로 텍스트 검색을 넘어 이미지를 이용한 멀티모달 검색이 가능해지면서 탐색의 범위가 확장되었고 자료 탐색 시간을 획기적으로 단축했다[1]. 그러나 여전히 아카이브 내 영상 데이터 관리는 여전히 전통적인 방식대로 사용자의 분류에 의존하고 개인화된 라벨링이 필요하다. 검색 과정이 개인이 직접 분류한 색인과 기억에 의존하므로 조건에 따라 검색 결과가 좌우된다는 단점과 함께, 비디오 양이 방대해지면서 색인 및 검토를 위한 인력 소모가 기하급수적으로 증가하고 있다. 이에 따라 저장된 데이터 내에서 효율적으로 비디오를 검색하기 위한 다양한 방법론이 제시되고 있다[2].

과거 인공지능을 이용한 영상 자료 검색 기술은 주로 추출된 메타 데이터 키워드에 의존하였다. 장면, 인물, 등장 객체 등에 해당하는 키워드와 장면 서술 문장을 사전에 추출하여 비교하는 형태로 사용자의 프롬프트와 유사한 내용의 콘텐츠 혹은 장면을 찾았다[3,4]. 이러한 선행 기술은 사전 추출 모델에 크게 의존하며 실제 방송 산업에서 활용되는 영상은 길이와 내용이 매우 다양하므로 키워드 기반의 검색 결과는 정확도가 떨어질 수 있다. 특히 드라마 및 방송영상의 경우 편집자의 의도에 따라 다양한 내용, 자료 화면 등이 포함되어, 짧은 영상 클립으로 학습하는 기존의 인공지능 모델이 효과적으로 적용되기 어렵다. 구체적으로 춤을 추는 사람, 운동을 하는 사람 등의 하나의 내용을 포함한 학습 데이터와 달리 실제 뉴스, 드라마, 촬영 영상 등의 방송영상은 인터뷰, 활동, 클로즈업, 자료 화면 등의 다양한 장면으로 구성되어 있다. 따라서 이러한 영상은 사람이 텍스트를 통해 라벨링 하기 어려우며 중복성을 제거하지 않고 모든 내용에 대한 메타 데이터를 취득하는 것 또한 많은 계산 및 저장 공간이 필요하다. 이를 해결하기 위해 본 제안 기술에서는 개인 아카이브를 효율적으로 검색할 수 있는 시스템을 구축하는 것을 목표한다.

본 논문에서는 생성형 모델과 멀티모달 사전학습 모델을 활용한 새로운 형태의 영상 검색 기술 방식을 제안하고 그 구체적인 제작 방식을 설명한다. 멀티모달 사전학습 모델[1]은 대용량 데이터를 통해 사전에 학습하여 영상과 텍스트 데이터를 같은 의미적 공간에서 상호 대응하여 사용할 수 있다. 이를 통해 텍스트와 유사한 의미적 맥락을 가진 영상 데이터를 쉽게 찾을 수 있으며, 검색의 질을 높이기 위해 생성형 이미지를 가이드로 하여 텍스트로 표현하지 못하는 장면의 구도 등의 추상적 개념을 구체화하여 검색에 사용한다.

이 같은 멀티모달 AI 검색 기술은 방송사가 보유한 방대한 양의 영상 아카이브를 효과적으로 활용할 수 있도록 해주어 방송 업계에 큰 혁신을 가져올 것으로 기대된다. 기존에는 수작업으로 아카이브 영상을 검색하고 관리해야 했기 때문에 많은 인력과 시간이 소요되었다. 하지만 이 기술을 통해 인물, 행동뿐만 아니라 추상적인 문구로도 영상을 검색할 수 있게 되어 방대한 양의 데이터를 취급하는 방송국에서 자료 검색과 관리의 효율성을 크게 높여줄 것으로 예상된다. 영상 AI 분석을 통한 메타 데이터 자동 생성 기술은 현재 세계 많은 미디어테크 기업들이 시도하고 있는 분야로, 자동 편집 기술의 기반이 될 것으로 예상된다. 라벨링 할 수 없는 불특정 자료화면 이미지까지 검색 가능한 멀티모달 AI 기술은 향후 자동 편집 기술을 위한 중요한 토대가 될 것으로 기대된다.

또, 정교한 검색 기능을 통해 시청자가 원하는 특정 장면이나 주제를 쉽게 찾아볼 수 있게 되면 영상의 활용도도 높아지고 방송사 간 콘텐츠 공유 및 거래도 용이해질 것으로 예상된다. 적확한 영상 검색과 제공이 가능해지면 콘텐츠 라이브러리 통합과 연계가 수월해지기 때문이다. 이 경우, 공적 가치를 지닌 공영 방송사의 아카이브 활용도도 높아지고 결과적으로 방송 산업의 다양한 측면에서 혁신적인 변화를 불러올 중요한 도구가 될 것으로 기대된다.

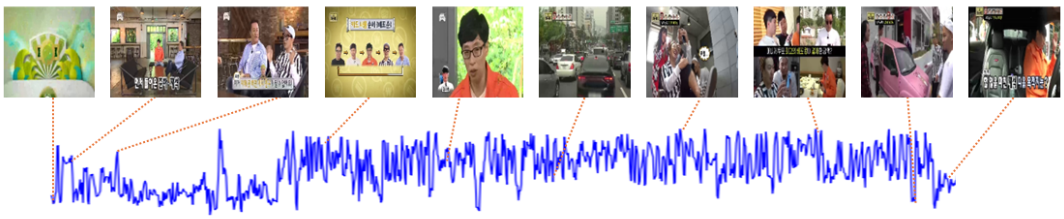
본 논문은 과학기술정보통신부와 한국전파진흥협회(RAPA)의 뉴테크 융합 제작실증 사업으로 진행하여 한국전파진흥협회에 제출한 최종 기술 보고서에 기반하여 작성한 것이다.

II. 관련 연구

1. 인공지능 기반 콘텐츠 및 장면 검색 기술

인공지능 기술을 이용한 장면 탐색은 주로 사전에 라벨링 된 텍스트와 타겟 문구 사이의 의미적 유사도를 검사하여 결과를 반환하는 방식으로 발전해 왔다. 대표적인 예로 한국전자통신연구원 (ETRI)에서는 객체, 배경 등의 메타데이터를 추출하고 해당 메타 데이터와 타겟 문장 간의 유사도를 검사하여 콘텐츠를 검색하는 시스템을 제안하였다[3]. 나아가 비디오 캡션 생성 모델을 이용하여 장면에 관한 서술을 생성하고 타겟 문장과 생성된 캡션 사이의 유사도 검색을 통한 콘텐츠 검색 시스템을 제안하였다[4]. 케이티(KT)에서는 장면별 키워드를 사전 추출하여 타겟 문장의 유사도를 검사하여 관련성이 높은 장면을 검색하는 방법을 제안하였다[5]. 그러나 이러한 시스템은 지도학습을 위해 정교하게 라벨링 된 데이터가 필요하였고, 키워드 추출의 성능은 사전에 학습된 객체 추출기 및 캡셔닝 모델에 의존하기 때문에 구체적인 맥락을 제공하지 못하고 새로운 상황에 대한 추론 능력이 부족하다는 한계를 갖는다. 더욱이, 텍스트 키워드의 일치도를 기반으로 하는 검색은 키워드에 따라 사용자 관점에서 정확도가 높지 않은 유효하지 않은 결과를 초래할 수 있다.

이러한 한계를 해결하기 위해 멀티모달 사전학습 모델을 이용한 검색 기술이 제안되었다. 멀티모달 사전학습 모델은 서로 다른 모달리티 간의 공동 표현을 학습하기 위해 트랜스포머[6]와 같은 새로운 신경망 구조, 대규모 데이터, 그리고 새로운 손실함수를 통해 빠른 발전과 성능향상을 이루었다[1,7,8]. 멀티모달 사전학습 모델은 학습한 공동 특징을 통해 텍스트-시각 검색, 캡션 생성, 질의응답 등 다양한 멀티모달 작업을 가능하게 하며 특히 서로 다른 모달리티 데이터 간의 유사도 비교를 통해 새로운 검색을 가능하게 한다. 비전-언어 멀티모달 사전학습[1,8]을 이용하면 대용량 이미지와 텍스트 페어 데이터를 활용하여 모달리티 간의 상호작용을 이해하고 이미지와 텍스트를 공동 특징 차원에 표현하는 것이 가능하다. 다시 말해, 멀티모달 사전학습 모델은 이미지와 텍스트 간의 상호작용을 이해하고 활용하는 모델로, 일반적이지 않은 상황 혹은 복잡한 묘사를 서술한



(a) Similarity score with target prompt : “A friendly people is gathering.”

(b) Top 5 selected image with target prompt (upper) and target prompt and generated image (lower)

그림 1. (a) 기존 의미적 유사도 기반 멀티모달 시멘틱 장면 검색 방법, (b) 기존 의미적 유사도 기반 장면 검색 방법과 제안하는 이미지 가이드 장면 검색 방법 비교. 기존 방법에 비해 다양하고 정교한 결과를 보임

Fig. 1. Comparison of (a) existing multi-modal semantic scene retrieval methods, (b) previous scene retrieval methods with our proposed generated image guided scene retrieval method. Results are diverse and sophisticated compared to existing methods.

이미지 검색에 적합하다.

대표적인 멀티모달 사전학습 모델인 CLIP[9]은 대용량 이미지와 텍스트 데이터 쌍을 활용하여 단일 모달 데이터를 인코딩하고, 대조학습을 통해 공동 엠베딩을 학습한다. 이미지 분류 태그로부터 문장 데이터를 생성하여, 과거의 모델들보다 정교한 라벨이 필요 없이 대용량 데이터로부터 공동 표현을 학습함으로써 처음 보는 이미지에 대해서도 강력한 성능을 보인다. CLIP은 이미지를 자연어로 설명하거나 텍스트와 이미지 간의 의미 관계를 파악하는 다양한 작업을 수행할 수 있는 다목적 모델로, 특히 CLIP을 이용한 텍스트 기반 장면 검색 기술에서의 성능을 크게 향상시켰다. CLIP4clip[10]은 앞서 제안된 CLIP 모델을 기반으로 한 연구로, 비디오를 대상으로 확장된 모델이다. 비디오 특징을 추출하기 위해 비디오에서 추출한 프레임을 2D 또는 3D 패치로 만들어 텍스트와 함께 학습하였다. 또한 비디오-텍스트 검색을 위해 Frozen[11]은 이중인코더 구조를 제안하여 이미지와 비디오 간의 유사도를 학습하며, 이때 이미지와 비디오 정보를 유연하게 학습시키기 위해 시간적 컨텍스트를 점진적으로 학습하는 커리큘럼 학습을 진행하였다. Bridgeformer[7]는 동사와 명사 즉, 주요 부분을 활용하는 질의 손실함수를 제안하여 비디오의 주요 영역에 대한 비디오-텍스트 사이의 의미론적 연관성을 학습하게 하고 비디오-텍스트 검색의 성능을 향상시켰다.

이처럼 멀티모달 모델을 통해 텍스트로부터 영상의 유사도를 측정할 수 있도록 하는 시멘틱 검색 기술이 높은 성능을 보임에 따라 실제 산업 현장에서도 실질적인 요구가 커지고 있다. ETRI[2]는 CLIP을 활용하여 텍스트, 이미지, 동영상을 활용한 멀티모달 사용자 질의를 지원하는 비디오 검색 기술 및 시스템을 제안했다. 그러나 그림 1의 (b)와 같이 실제 방송 영상에 사용할 때 모델이 처음 보는 화면에 대한 추론 능력의 한계와 타겟 문장과 일치하는 장면이 중복되어 나타나는 상황에서 검색 효율이 저하되는 문제가 있었다. 특히 영상은 다양한 스타일의 장면을 담고 있지만 검색하고자 하는 내용이 다수의 검색 결과와 유사한 경우 텍스트 유사도를 바탕으로 한 결과는 사용자의 의도에 맞지 않는 경우가 많다. 텍스트로 표현할 수 없는 화면의 구도 분위기 등의 추상적인 개념이 영상에 포함되기 때문이다.

기존기술과 차별화하여 제안 기술에서는 생성형 모델을 통한 가이드 이미지를 생성하여 이러한 한계를 극복하였다. 그림 1, 2와 같이 생성형 이미지는 텍스트로 구체화할 수 없는 장면의 구도, 구체적 묘사 등을 프롬프트를 통해 지시하고 생성하기 적합하다. 따라서 본 기술에서는 생성형 가

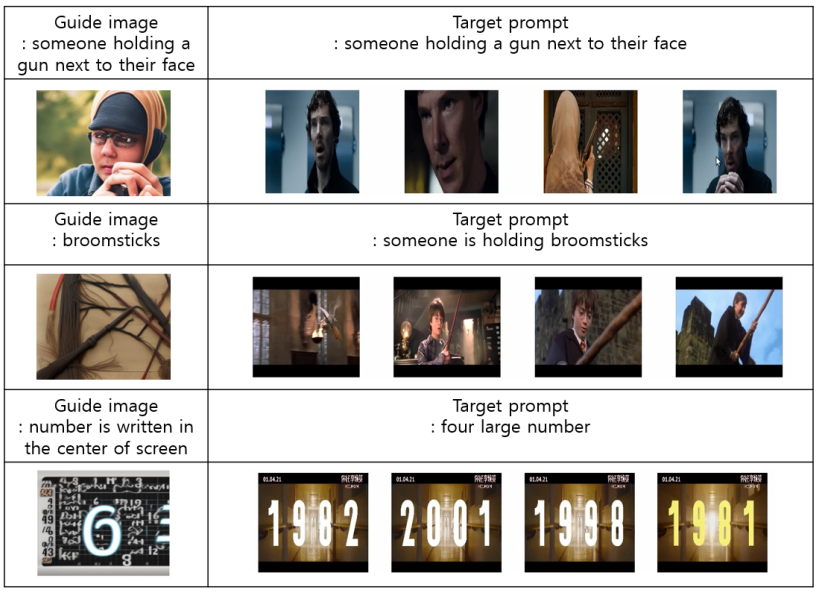


그림 2. 이미지 가이드를 이용한 콘텐츠 검색 예시
Fig. 2. Content search with image guide and text prompt

이드를 이용한 멀티모달 검색 기술의 제안을 통해 사용자가 간단하게 텍스트 입력을 통해 원하는 컨텐츠 및 장면을 빠르게 찾는 방식을 제안한다.

2. 이미지 생성 모델

생성형 인공지능은 사용자의 요구사항에 따라 새로운 데이터를 생성하는 인공지능으로 최근 많은 미디어 산업에서 생성형 인공지능의 활용이 늘고 있다. 대표적인 예로 Stable Diffusion[12]은 Diffusion 모델 기반의 생성형 모델로, 사용자가 입력한 프롬프트에 대한 이미지를 생성하는 기능을 제공한다. Diffusion Model은 노이즈가 추가된 이미지에서 노이즈를 제거하는 디노이징 과정을 학습해 이미지를 생성하는 모델이다. Stable Diffusion은 사용자가 입력한 텍스트 설명에 따라 상세한 부분까지 생성 이미지를 조절할 수 있고 다양한 스타일의 이미지를 생성할 수 있다. 또한 오픈소스로 공개되어 Stable Diffusion을 기반으로 하는 다양한 AI 이미지 서비스들이 생겨나고 있다. 이를 이용하여 비디오를 생성하는 연구도 진행되고 있으며 새로운 컨텐츠 생성 기법으로 주목을 받고 있다. 본 기술에서는 사용자의 의도라는 추상적인 개념을 이미지로 구체화함으로 검색에서 검색 결과의 개인화를 위한 지능형 비디오 검색 시스템을 제안한다.

Ⅲ. 멀티모달 장면 검색 시스템

멀티모달 사전학습 모델과 이미지 생성 모델을 조합하여, 텍스트 프롬프트를 통한 지능형 비디오 검색 시스템을 설명한다. 제안하는 장면 검색 기술은 방송 아카이브 내에서 효과적으로 영상을 검색할 수 있도록 비디오 특징 추출, 이미지 생성, 멀티모달 검색 모듈로 구성되며 검색 단계에서의 동작 방식은 그림 3과 같다. 사용자는 시스템을 통해 텍스트 혹은 텍스트와 이미지를 모두 사용하여 원하는 결과를 얻을 때까지 반복하여 장면 탐색을 할 수 있다. 그림 3에서와 같이, 이미지 가이드를 필요로 하는 모드 혹은 그렇지 않은 모드 중 하나를 선택하고 원하는 장면 혹은 콘텐츠에 관해 서술하는 타겟 프롬프트를 입력한다.

이후 필요시 이미지 생성을 위한 프롬프트를 작성하여 원하는 이미지를 생성한다. 생성된 이미지는 사용자의 의도를 구체화한 결과로, 현실에서 존재하기 힘든 이미지까지도 생성하여 검색의 범위를 확장한다. 사용자는 이미지 생성 단계에서 추가적인 프롬프트 수정과 이미지 재생성을 통해 의도에 부합하는 정확한 이미지 가이드를 획득할 수 있다. 생성 이미지와 사용자가 입력한 타겟 프롬프트는 공동 엠베딩으로 표현되고 아카이브 내의 컨텐츠 특징과 비교하여 가장 연관도가 높은 비디오를 사용자에게 반환한다.

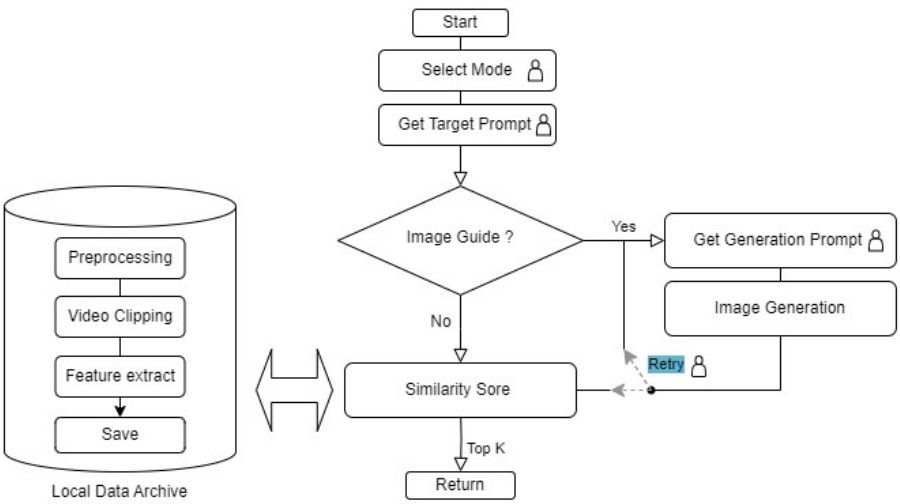


그림 3. 제안하는 비디오 검색 시스템 동작 방식
Fig. 3. Proposed video retrieval system workflow

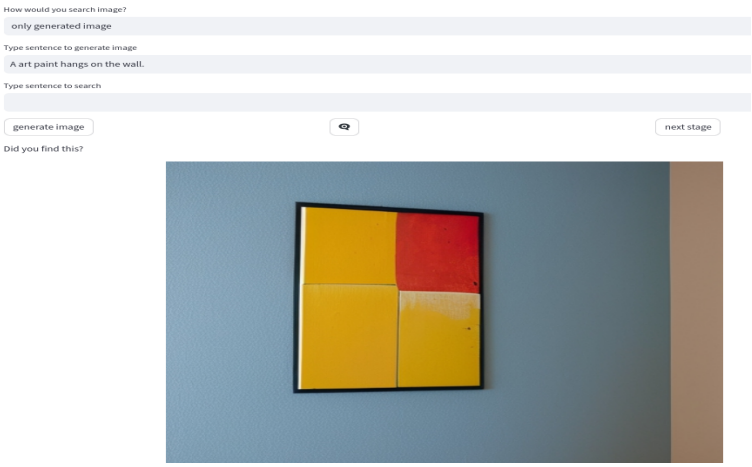


그림 4. 생성된 이미지 예시 “an art paint hangs on the wall”
Fig. 4. Example of generated image with prompt of “an art paint hangs on the wall”

1. 비디오 특징 추출부

본 시스템은 비디오 특징을 추출하기 위해 사전학습 된 멀티모달 모델 Frozen[11]을 활용한다. 해당 모델은 비디오 엔코더와 텍스트 엔코더로 구성된 이중 스트림 구조를 가지며 대용량 이미지-텍스트 쌍을 통해 학습되었다. 로컬 내 컨텐츠는 Frozen의 비디오 엔코더를 통해 비디오 특징을 추출한다. 다양한 길이의 비디오를 효과적으로 저장하고 관리하기 위해 5분 이하의 짧은 비디오는 초 단위의 짧은 영상 클립으로 분할되며, 5분 이상의 장기 비디오는 Python 기반의 장면전환 탐색 패키지 PySceneDetect을 이용하여 장면 단위로 비디오 특징을 추출한다. 저장된 영상 특징은 로컬 내 아카이브에 저장되며 검색 단계에서 접근된다.

2. 이미지 생성부

이미지 생성을 위해 stability-AI에서 공개한 Stable Diffusion 2[12] 모델을 사용하였다. Stable Diffusion 2 모델은 최신의 딥러닝 기반 이미지 생성 모델 중 하나로, CLIP 모델을 텍스트 인코더로 사용하여 제공한 텍스트 설명을 바탕으로 상세하고 현실적인 이미지를 생성한다. 생성 이미지는 사용자가 입력한 프롬프트를 기반으로 생성되며 사

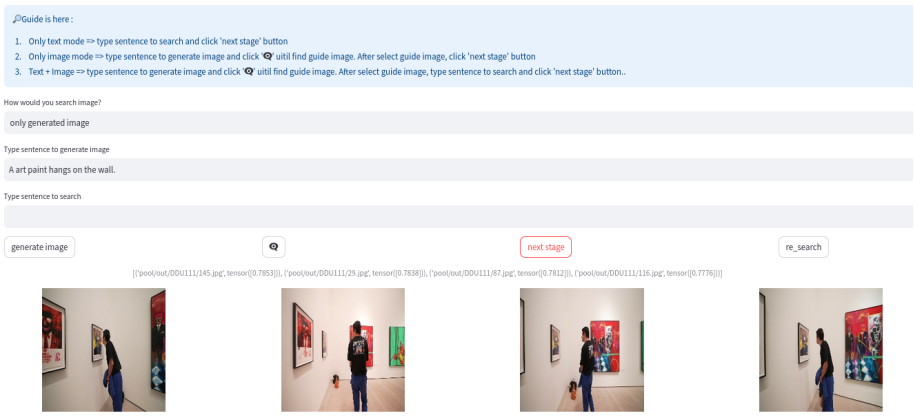


그림 5. 생성된 이미지 그림 4를 이용한 멀티모달 검색 결과
Fig. 5. Results of multi-modal video search with guided image

용자는 이미지 생성 단계에서 원하는 이미지 가이드가 나올 때까지 반복하여 생성하여 타겟 이미지를 얻는다. 선택된 가이드 이미지는 멀티모달 검색부에서 중첩되어 비디오 특징으로 엔코딩되고, 검색 단계에서 사전 추출된 콘텐츠 엠베딩과의 유사도 비교를 통해 이미지 유사도 기준을 결정한다. 같은 이미지 생성 프롬프트에 대해 다양한 이미지가 생성될 수 있으며, 프롬프트의 구체화 수준에 따라 생성된 이미지는 큰 차이를 보인다. 이는 노이즈로부터 텍스트 프롬프트에 맞게 이미지를 생성하는 생성형 모델의 특징이며, 제안 방법은 이러한 특징으로부터 아이디어를 얻어, 다양한 이미지를 생성하고 아카이브 내에서 다양한 구체적이면서도 다양한 컨텐츠 검색을 가져오는 것을 목표로 한다. 그림 4는 제작 소프트웨어 (SW)의 사용자 인터페이스 (UI)를 통해 입력한 텍스트와 생성된 이미지를 보인다.

3. 멀티모달 검색부

멀티모달 검색을 위해 입력된 타겟 프롬프트는 텍스트 엔코더를 통해 텍스트 엠베딩으로 출력된다. 텍스트 벡터는 입력 이미지와 함께 코사인 유사도를 통한 유사도 비교를 진행하여 가장 관련도 높은 결과를 출력한다. 생성된 이미지와 측정된 유사도 기준과 텍스트 유사도를 최종 검색 결과는 생성된 이미지, 타겟 프롬프트 그리고 아카이브 콘텐츠 사이의 유사도 비교를 통해 반환된다. 이미지 가이드를 선택적으로 반영하기 위해 최종 유사도 점수 s는 이미지 유사도와 텍스트 유사도의 가중합으로 표현된다. 0~1 사이의 가중치 파라미터를 통해 이미지 유사도 기준을 강조하거나 텍스트 유사도를 강조하여 사용한다. 그림 5는 제작 소프트웨어 (SW)의 사용자 인터페이스 (UI)를 통해 생성된 이미지를 이용한 멀티모달 검색 결과이다.

Ⅳ. 실험

성능 평가를 위해 비디오 리트리벌 학습 및 평가데이터로 사용되는 1만개의 비디오-텍스트 쌍으로 구성된 MSR-VTT[13] 데이터 셋과 방송사 아카이브로부터 선별한 다양한 카테고리의 100개의 영상을 이용하였다. MSR-VTT에 대한 “text to video retrieval” 태스크를 통해 검색 시스템의 성능을 평가한다. 평가지표인 재현률 (Recall @ K) 평가지표는 K개의 추천 결과에 대해 연관된 결과가 모델의 추천 결과에 얼마나 포함되는지를 의미한다. 평가 결과는 아래의 표 1에서와 같다. 표 1에서 베이스라인 모델로 채택한 Frozen과 최신 멀티모달 사전학습 모델인 VIOLETv2, Bridgeformer과 생성형 가이드를 추가한 제안 방법의 결과를 비교한다. 생성형 이미지는 리트리벌을 위한 타겟 쿼리를 통해 사전에 생성하였으며, 이미지 가이드를 위한 가중치 파라미터는 0.2로 설정되었다. 표에서 보이듯이 베이스라인 모델 Frozen과의 재현률 대비 성능이 Recall @ 1에서는 7.5, Recall @ 5에서는 9.6, Recall @ 10에서는 11.7 향상된 것을 확인하였다.

표 1. MSR-VTT데이터 셋을 이용한 Text-Video Retrieval 태스크 결과
Table 1. Result of Text-Video Retrieval on MSR-VTT dataset

	Recall @ 1	Recall @ 5	Recall @ 10
Frozen[11]	31.0	59.5	70.5
Bridgeformer[7]	37.6	64.8	75.1
VIOLETv2[8]	37.2	64.8	75.8
Ours	38.5	69.1	82.2

또한 방송사 아카이브로부터 선별한 100개의 테스트한 영상은 사람이 사전에 설정한 키워드를 이용하여 성능 평가를 진행하였다. 키워드는 “눈 오는 날 거리”, “가족 식사” 등과 같이 짧은 레이블이며, 타겟 문장은 “The video of <key word>”로 설정한 뒤 검색 결과의 정확도(accuracy)를 측정하였다. 표 2에서 보이듯이 정확도가 16.1% 향상된 것을 확인하였다.

표 2. 아카이브 영상 검색 성능 비교
Table 2. Comparison on archive video data

	Accuracy (%)
Frozen[11]	70.6
Ours	86.7

또한 연구진은 문화방송(MBC) 방송 아카이브를 활용해 이러한 검색 모델의 성능을 검증하였다. 특히 장시간의 예능 방송 영상을 사용하여 정성적 평가를 진행하였다. 구체적으로 실제 편집물 생성자가 생성형 AI 모델을 활용해

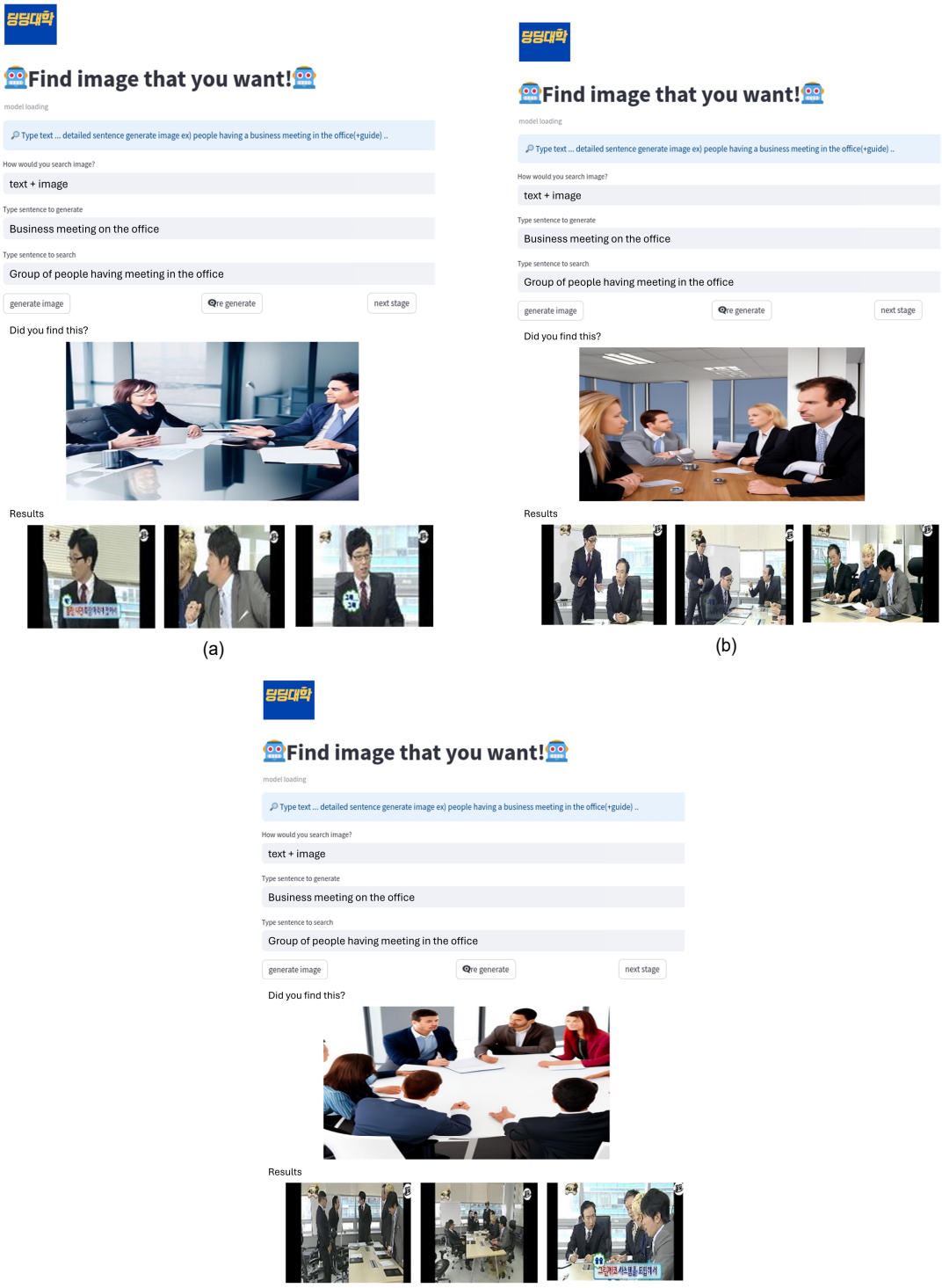


그림 6. 무한도전 영상 검색 결과 예시. 프롬프트 구체화 과정을 통해 생성 이미지에 (a) 1~2 명의 등장인물 (b) 3~4명의 등장인물 (c) 그 이상의 등장인물이 나타나는 경우의 이미지를 생성하고 동일 타겟 쿼리에 대한 비디오 장면 검색 결과

Fig. 6. Results of scene detection in “infinite challenge”, when using different style of generated image

사람들이 모여 회의하는 장면을 생성한 뒤, 이와 유사한 장면을 <무한도전>, <라디오스타>, <놀면 뭐하니> 등의 예능 프로그램에서 검색하였다. 방송 영상물을 제작할 때는 주로, 특정 화면뿐 아니라 불특정한 자료 화면을 검색하는 경우가 많고, 특히 예능 프로그램에서 재미를 주기 위한 짤막한 삽입 영상이 들어가는 경우가 많은데, 텍스트 만을 사용하여 검색하였을 때는 비디오의 중복된 장면을 다수 보유하는 특징으로 인해 일관된 중복이미지를 계속하여 반환하는 경향을 보였다. 그러나 생성형 이미지를 가이드로한 검색 결과는 그림 6과 같이 다양하고 정확한 결과를 보이며 이런 경우에 높은 활용 가능성을 보였다. 그림 6은 “Buisniss meeting on the office”라는 프롬프트를 기반으로 장면을 생성하고, 약 1시간 길이의 무한도전 방영 비디오로부터 “group of people having meeting in the office”라는 타겟 쿼리에 대해 장면 검색을 시행한 결과이다. 테스트 단계에서 프롬프트 수정을 통해 등장인물의 수, 구도 등을 구체화하여 그림 6의 (a),(b),(c)와 같은 다른 스타일의 회의 장면을 생성하였다. 같은 타겟 쿼리에 대해 이미지 가이드를 따라 컨텐츠 내의 다른 장면을 반환하는 것을 확인할 수 있다.

또한 다양한 검색 결과 예시를 통해 제안하는 검색 시스템의 정성적 평가를 진행하였다. 그림 7에서 이미지 가이드가 있을 때와 그렇지 않은 상황에 대한 검색 결과를 비교하였다. 이미지 가이드가 텍스트로 표현하기 힘든 특징을 제공하여 사용자의 의도에 적합한 결과를 반환하는 것을 확인할 수 있다.

V. 결론 및 한계

본 연구에서는 생성형 인공지능 모델과 멀티모달 사전학습 모델을 융합한 새로운 영상 검색 시스템을 제안하였다. 이 시스템은 사용자의 텍스트 프롬프트와 생성된 가이드 이미지를 활용하여 방송사의 방대한 아카이브로부터 정확하고 효율적인 영상 검색을 가능하게 한다.

제안된 기술은 크게 세 가지 주요 모듈로 구성된다. 첫째, 비디오 특징 추출부에서는 사전학습된 멀티모달 모델인 Frozen을 활용하여 아카이브 내 비디오의 특징을 추출한다. 둘째, 이미지 생성부에서는 Stable Diffusion2 모델을 사용하여 사용자의 텍스트 프롬프트에 따라 가이드 이미지를 생성한다. 셋째, 멀티모달 검색부에서는 생성된 이미지와 텍스트 프롬프트를 공동 엠베딩 공간에 매핑하고, 비디오 특징과의 유사도를 계산하여 최종 검색 결과를 제공한다.

제안 기술의 성능을 평가하기 위해 MSR-VTT 데이터셋과 실제 방송사 아카이브 영상을 활용하였다. 실험 결과, 기존 방법 대비 Recall@1에서 7.5%, Recall@5에서 9.6%, Recall@10에서 11.7%의 성능향상을 이뤘으며, 아카이브 영상 검색에서도 정확도가 16.1% 향상되었다. 또한 다양한

그림 7. 텍스트 입력만 사용한 경우와 가이드 이미지를 사용한 장면 검색 비교
Fig. 7. Comparison between text-only input and scene search using a guide image

검색 결과 예시를 통해 제안 기술이 텍스트로 표현하기 힘든 특징을 효과적으로 반영하여 사용자 의도에 부합하는 결과를 제공함을 확인하였다.

본 기술은 사용자가 로컬 환경에서 보유한 데이터에 대한 보다 효율적인 검색을 가능하게 한다. 사용자는 간단한 이미지 생성 프롬프트와 타겟 텍스트 입력만으로 원하는 컨텐츠 및 장면을 신속하게 검색할 수 있다. 이를 통해 기존의 수작업 기반 검색 과정에서 소요되던 시간과 노력을 획기적으로 절감할 수 있으며, 실제 방송 영상을 활용한 실험에서도 이와 같은 결과를 확인할 수 있었다. 제안 기술의 활용 가능성은 매우 크다. 대형 방송 제작사는 물론 1인 미디어 제작자들도 쉽게 사용할 수 있는 기술로, 방송 콘텐츠 제작 및 아카이브 탐색 과정에서 시간과 자원을 절약하는 데 크게 기여할 것으로 기대된다. 또한 이 기술은 불특정 화면의 검색을 용이하게 해, 궁극적으로 자동 편집 기술의 기반이 되어 방송 생태계 전반의 혁신과 발전을 가져올 것으로 전망된다.

그러나 Stable Diffusion을 통해서 얻은 생성 이미지는 사용자의 의도에 부합하기 위해 보장되지 않은 반복 생성 혹은 프롬프트 수정이 필요하다. 또한 프롬프트를 통해 구체적인 지시를 내리는 것에는 해당 모델에 대한 경험을 통해 획득한 통찰을 필요로 한다. 이와 같은 한계로 인해 프롬프트에 대한 자세한 안내 없이는 제안 검색 시스템은 활용하기 어렵다는 한계가 존재한다.

향후 연구로는 대규모 아카이브 데이터에 대한 실증 적용을 통해 기술의 안정성과 확장성을 제고할 필요가 있다. 또한 생성 이미지의 품질 개선과 가이드 이미지 선택 기준 최적화를 통해 검색 정확도를 더욱 향상할 수 있을 것으로 기대된다.

참 고 문 헌 (References)

[1] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu,“Coca: Contrastive captioners are image-text foundation models,” arXiv preprint arXiv:2205.01917, 2022.
doi: https://doi.org/10.48550/arXiv.2205.01917

[2] W. Park and N. Lee, “A study on Multimodal Video Retrieval using the CLIP”, Korea Artificial Intelligence Conference, Jreju, Korea, pp.109-110, 2022.

[3] ETRI, Method for generating metadata based on scene representation using vector and apparatus using the same, KR20210122496A, Korea, 2021.

[4] KT, System and method for generating keyword information from each moving picture scene, KR20110014403A, Korea, 2011.

[5] ETRI, Apparatus and method for generating visual annotation based on visual language, KR102156440B1, Korea, 2021.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, 30, Long Beach, CA, USA, 2017.
doi: https://doi.org/10.48550/arXiv.1706.03762

[7] G. Yuying, et al., "Bridging video-text retrieval with multiple choice questions," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana, USA, 2022.
doi: https://doi.org/10.48550/arXiv.2201.04850

[8] Fu, Tsu-Jui, et al. "An empirical study of end-to-end video-language transformers with masked visual modeling." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023.
doi: https://doi.org/10.48550/arXiv.2209.01540

[9] R, Alec, et al., "Learning transferable visual models from natural language supervision," International conference on machine learning, Virtual Only, 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020

[10] H. Fang, P. Xiong, L. Xu and Y. Chen, “Clip2video: Mastering video-text retrieval via image clip,” arXiv preprint arXiv:2106.11097, 2021.
doi: https://doi.org/10.48550/arXiv.2106.110

[11] M. Bain, et al., "Frozen in time: A joint video and image encoder for end-to-end retrieval." Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
doi: https://doi.org/10.48550/arXiv.2104.00650

[12] R. Rombach, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Virtual Only, 2022.
doi: https://doi.org/10.48550/arXiv.2112.10752

[13] J. Xu, et al. "Msr-vtt: A large video description dataset for bridging video and language." Proceedings of the IEEE conference on computer vision and pattern recognition. Las vegas, USA, 2016.

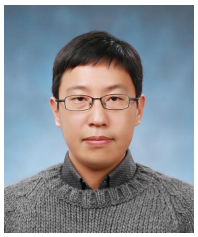
저 자 소 개

이 주 희

- 2020년 2월 ~ 현재 : Ewha Womans University, Department of Electronic and Electrical Engineering
- ORCID : https://orcid.org/0000-0002-9245-4688
- 주관심분야 : Video-Language pre-training / Video, language, and reasoning

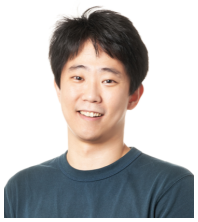
양 효 걸

- MBC AI 전략자회사준비TF장
- 다변수 최적화 및 시계열 분석
- 데이터 기반 인과 추론
- 데이터의 통계적 분석 및 시각화
- ORCID : https://orcid.org/0009-0002-6401-0614
- 주관심분야 : 아카이브 인덱싱 및 활용



강 제 원

- 2008년 8월 ~ 2012년 7월 : Ph.D, University of Southern California
- 2012년 8월 ~ 2014년 2월 : Senior engineer, Qualcomm Inc.
- 2014년 3월 ~ 현재 : Professor, Ewha Womans University
- ORCID : https://orcid.org/0000-0002-1637-9479
- 주관심분야 : Multimodal AI model, Video coding and processing



한 치 영

- MBC AI 전략자회사준비TF 기술책임
- IT 서비스 PM/TPM
- 데이터 시각화, 플랫폼 엔지니어링/DevOps
- ORCID : https://orcid.org/0009-0006-9095-3818
- 주관심분야 : ML기술의 대규모 처리 및 산업화

염 규 현

- MBC AI 전략자회사준비TF
- AI 아카이브 검색 UI/UX 기획 및 설계
- 생성형 AI를 활용한 아카이브 메타데이터 AX 연구
- ORCID : https://orcid.org/0009-0004-6243-1284
- 주관심분야 : 메타데이터 처리, AI 검색, AI 기술 국제 규제 및 저작권 이슈