논문 2025-2-7 http://dx.doi.org/10.29056/jsav.2025.06.07

멀티모달 행동 및 음성 인식 기반의 OTT 콘텐츠 장면 검색 고도화 방법

박승현*, 유인재*, 박병찬*, 김석윤*, 김영모*†

An Enhanced Scene Retrieval Method for OTT Content Based on Multimodal Action and Speech Recognition

Seung-Hyeon Park*, In-Jae Yoo*, Byeong-Chan Park*, Seok-Yoon Kim*, Young-Mo Kim*†

요 약

OTT(Over-the-Top) 플랫폼의 확산으로 인해 영상 콘텐츠의 양이 급증함에 따라, 정밀한 의미 중심적인 장면 검색 기술의 수요가 높아지고 있다. 특히 드라마나 영화와 같이 인물 중심의 서사와 다양한 편집 기법이 적용된 콘텐츠에서는 기존의 키프레임 기반 검색 방식만으로는 장면 간의 맥락과 의미를 효과적으로 파악하기 어렵다. 본 논문에서는 Transformer 기반의 행동 인식 모델과 STT(Speech-to-Text) 기반의 음성 인식 기술을 결합한 멀티모달 행동 및 음성 인식 기반의 OTT 콘텐츠 장면 검색 고도화 방법을 제안한다. 제안된 방법은 연속된 프레임을 의미 단위의 행동 구간으로 분할하고, 구간 내 객체 및 관계 정보를 통합하여 중복 없는 씬 그래프를 구성한다. 여기에 발화 구간을 정밀하게 추출하여 시각 정보와 시간적 정합성을 갖춘 형태로 통합함으로써, 기존보다 더 정밀하고 맥락 중심적인 장면 검색을 실현한다. 본 논문에서는 인물 중심 장면 탐색, 감정 기반 클립 추출, 발화 중심 검색 등 다양한 응용 가능성을 제시하며, 향후 OTT 콘텐츠의 활용성과 개인화된 콘텐츠 경험을 크게 향상시킬 수 있을 것으로 기대된다.

Abstract

With the rapid growth of OTT (Over-the-Top) platforms, the volume of video content has increased exponentially, leading to a rising demand for precise and context-aware scene retrieval technologies. In narrative-driven content such as dramas and films, which often involve complex editing techniques and character-centric storytelling, conventional keyframe-based search methods fall short in capturing semantic continuity and scene context. This paper proposes an advanced method for OTT content scene retrieval based on multimodal action and speech recognition, combining a Transformer-based action recognition model with Speech-to-Text (STT) technology. The proposed approach segments continuous video frames into meaningful action intervals and constructs a de-duplicated scene graph by integrating key objects and their relationships within each segment. Furthermore, speech segments are accurately extracted and temporally aligned with visual data, enabling a unified multimodal representation of scenes. This integration supports more refined and semantically rich scene searches, such as character-centered navigation, emotion-based clip extraction, and dialogue-driven retrieval. The proposed method is expected to significantly enhance the personalization and reusability of OTT content in various user-centered applications.

한글키워드 : 멀티모달 장면 검색, Transformer 기반 행동 인식, 음성-시각 정보 통합, OTT 콘텐츠 분석, 의미 중심 씬 그래프

keywords : Multimodal Scene Retrieval, Transformer-based Action Recognition, Speech-Visual Information Integration, OTT Content Analysis, Semantics-Centered Scene Graph

* 숭실대학교 컴퓨터학과
† 교신저자: 김영모(email: ymkim828@ssu.ac.kr)

접수일자: 2025.05.02. 심사완료: 2025.05.22.
게재확정: 2025.06.20.

1. 서 론

OTT(Over-The-Top) 플랫폼의 급격한 성장으로 영상 콘텐츠의 소비와 생산이 폭발적으로 증가하고 있다[1-2]. 이는 시청자의 콘텐츠 접근성을 높이고 미디어 활용 방식에 혁신을 가져왔지만, 동시에 방대한 영상 데이터를 효율적으로 탐색하고 활용하기 위한 기술적 요구를 가속화하고 있다[3-4]. 예를 들어, KBS 아카이브에 따르면 현재까지 누적된 아카이빙 콘텐츠는 100만 시간을 초과하며, 이 수치는 지속적으로 증가하고 있다[5]. 이러한 환경에서는 정교하고 지능적인 콘텐츠 검색 방법이 필수적이다.

특히 드라마나 영화와 같이 인물 중심의 서사를 갖는 콘텐츠는 장면 간의 시각적 전환이 빈번하고, 하나의 사건이 다양한 촬영 각도나 구도를 통해 표현된다[6]. 이로 인해 동일한 의미를 가진 장면이라도 시각적으로 상이하게 나타나며, 단순한 프레임 유사성이나 메타데이터에 기반한 기존 검색 방법만으로는 콘텐츠의 의미적 흐름을 정확히 파악하기 어렵다[7-8]. 또한 시청자의 검색 의도는 단순한 장면 이미지가 아니라 장면의 맥락, 인물의 행동, 발화 내용 등을 포함한 '의미 기반 탐색'에 가까워지고 있다[9].

이에 본 논문에서는 멀티모달 행동 및 음성 인식 기반의 OTT 콘텐츠 장면 검색 고도화 방법을 제안한다. 본 연구는 Transformer 기반의 행동 인식 모델을 활용하여 연속된 프레임을 의미 단위의 '행동 구간'으로 분할하고, 해당 구간에서 핵심 객체 및 관계를 추출하여 중복 없는 씬 그래프(Scene Graph)를 구성한다. 이후, STT(Speech-to-Text) 기반 음성 인식을 통해 발화 구간을 식별하고, 이를 시각적 정보와 통합함으로써 장면의 의미적 맥락을 고도화한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 분석 및 행동 인식, 음성 인식 기반 발화 분석, 멀티모달 통합 검색 방법을 기술한다. 3장에서는 본 논문에서 제안하는 방법인 멀티모달 행동 및 음성 인식 기반 장면 검색 고도화 방법을 기술한다. 4장에서는 본 논문에서 제안하는 방법을 이용한 멀티모달 장면 검색 기술의 응용 시나리오를 기술하며, 5장에서 결론으로 마무리한다.

2. 관련 연구

2.1 시각 기반 장면 분석 및 행동 인식

기존 장면 검색 기술은 주로 CNN 기반의 특징 추출과 키프레임 유사도 분석을 활용해 왔다. 하지만 이러한 방법은 장면 간의 시간적 흐름이나 의미 연결성을 반영하기 어려워, 맥락 중심 검색에는 한계가 있다. 이를 보완하기 위해 최근에는 Transformer 기반 행동 인식 모델이 도입되었으며, Kinetics, UCF101, HMDB51 등 공개 데이터셋을 활용한 시계열 기반 인간 행동 인식 연구가 활발히 진행되고 있다. 이러한 모델은 단순한 프레임 단위의 특징이 아닌, 시간 축에서의 연속적인 행동 패턴을 인식할 수 있다는 강점을 가진다[10-11].

2.2 음성 인식(STT) 기반 발화 분석

영상 내 발화 내용을 텍스트로 변환하기 위한 STT(Speech-to-Text) 기술은 음향 전처리, 음소 추출, 언어 모델 기반 문장 예측 등 복합적 처리 과정을 포함한다. 기존의 전통적인 음성 인식 방식은 억양, 방언, 잡음 등 환경 변수에 취약하였으나, 최근의 Whisper, Conformer, DeepSpeech 등의 딥러닝 기반 STT 모델은 다양한 상황에서도 안정적인 텍스트 변환을 지원하고 있다. 이로 인해 장면 내 대사 기반 콘텐츠 탐색이 점차 실용화되고 있다[12-13].

2.3 멀티모달 통합 검색 기법

시각 정보와 음성 정보를 통합하여 장면을 정밀하게 인식하려는 멀티모달 검색 기술 또한 주목받고 있다. 대표적으로 이미지-텍스트-오디오 기반 씬 그래프(Scene Graph)를 구축해 콘텐츠를 구조화하고, 이를 통해 복합 질의에 대응하려는 접근이 제안되고 있다. 그러나 대부분의 기존 연구는 짧은 클립이나 단일 이벤트 중심 분석에 국한되며, 방송·영화와 같은 다각도 편집, 감정 변화, 서사 전개가 복합적으로 얽힌 장면을 정밀하게 검색하는 데는 한계가 있다[14].

3. 멀티모달 행동 및 음성 인식 기반 장면 검색 고도화 방법

3.1 개요

본 논문에서 제안하는 OTT 콘텐츠 장면 검색 고도화 방법은 행동 인식 기반 시각 정보 분석과 음성 인식(STT) 기반 텍스트 정보를 통합하는 멀티모달 프레임워크로 구성되며, 그림 1과 같다.

이를 통해 장면의 시각적 연속성과 대화 맥락을 동시에 반영할 수 있는 고도화된 검색이 가능하다. 전체 구성은 행동 구간 분할, 씬 그래프 생성 및 통합, 음성 인식 및 발화 구간 추출, 멀티모달 정보 통합 네 단계로 구성된다.

3.2 행동 인식 기반 시각 정보 분석

OTT 콘텐츠는 다양한 인물의 동작과 복잡한 장면 전환으로 구성되어 있으며, 단일 프레임 수준의 분석만으로는 장면의 의미를 정밀하게 파악하기 어렵다. 따라서 본 논문에서는 Transformer 기반의 행동 인식 모델을 활용하여, 연속된 프레임을 의미 단위의 행동 구간으로 분할한다.

입력 영상은 시간 축에 따라 순차적인 프레임 $V = \{F_1, F_2, \ldots, F_T\}$로 구성되며, 각 프레임에 대해 행동 인식 모델 M은 식(1)과 같이 행동 클래스를 예측한다.

$$M(F_t) \rightarrow a_k \in A, where\ A = \{a_1, a_2, \ldots, a_k\} \quad (1)$$

동일한 행동 a_k가 일정 시간 동안 지속될 경우, 이를 행동 구간 S_i인 식(2)로 정의할 수 있다.

$$S_i = \{F_t | M(F_t) = a_k, s_i \leq t \leq e_i\} \quad (2)$$

이러한 구간은 시각적으로 일관된 의미를 가지는 단위로 간주되며, 이후의 객체 및 관계 분석에 사용된다.

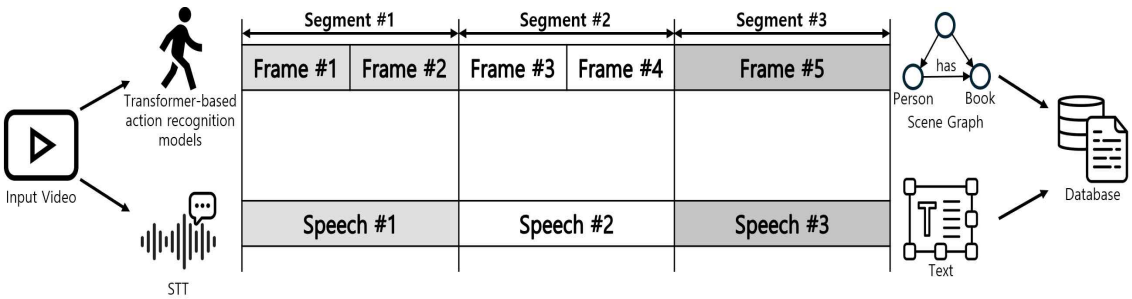

그림 1. 멀티모달 콘텐츠 분석을 위한 세그먼트 기반 처리 구조
Figure 1. Segment-Based Processing Architecture for Multimodal Content Analysis

각 행동 구간에서 등장하는 주요 객체(예: 인물, 물체)와 이들 간의 관계(예: 손잡기, 바라보기 등)를 추출하여 씬 그래프(Scene Graph)를 구성한다. 씬 그래프는 객체를 노드로, 관계를 엣지로 표현하는 그래프 구조로, 장면의 의미 구조를 시각적으로 표현할 수 있다.

그러나 연속된 프레임마다 별도의 그래프를 생성할 경우 동일 객체 및 관계가 반복되어 정보가 중복되고 검색 효율이 저하된다. 이에 따라 본 연구에서는 구간 내 프레임들로부터 핵심 관계만을 추출하고, 의미적으로 통합된 형태의 그래프를 생성하며, 표 1은 생성되는 예시이다.

또한, 통합된 씬 그래프의 예시는 다음과 같이 시각화될 수 있다.

- Frame A: Person → hold → Book
- Frame B: Person → read → Book

통합 그래프: Person → {hold, read} → Book

표 1. 행동 구간별 인식 결과
Table 1. Recognition results by action segment

행동 구간	프레임 범위	인식된 행동	주요 객체	주요 관계
S_1	F_1 ~ F_4	앉기	Person, Chair	sit-on
S_2	F_5 ~ F_8	책 읽기	Person, Book	hold, read
S_3	F_9 ~ F_{11}	고개 끄덕이기	Person	nod

3.3 음성 인식 기반 발화 정보 처리

영상 콘텐츠에서 음성 정보는 인물 간의 상호작용, 감정 표현, 그리고 내러티브 전개를 이해하는 데 있어 중요한 단서를 제공한다. 특히 OTT 콘텐츠와 같이 대사 중심의 서사가 전개되는 영상에서는 발화 내용을 정밀하게 분석하는 것이 장면 의미 파악에 결정적이다. 이에 본 연구는 음성 정보를 효과적으로 처리하기 위해 STT(Speech-to-Text) 기반 기술을 활용한다. STT 기반의 발화 정보 처리는 음향 기반 발화 구간 탐지, STT 변환 및 자연어 텍스트 생성 그리고 행동 구간과의 정합성 통합 세 단계로 구성된다.

첫 번째 단계인 음향 기반 발화 구간 탐지에서는 먼저 입력된 영상의 오디오 트랙에 대해 음향 신호 분석을 수행하여, 특정 시간 구간에서 발화가 발생하는지를 판단한다. 이 과정에서는 음성 주파수, 에너지, 피치(pitch) 등의 특징을 분석하여 발화(voice activity)와 비발화(non-speech) 구간을 구분한다. 이를 통해 정확한 발화 구간 $T_S = [t_{start}, t_{end}]$을 정의할 수 있다.

두 번째 단계인 STT 변환 및 자연어 텍스트 생성에서는 식별된 발화 구간에 대해 STT 모델을 적용하여 음성을 자연어 문장으로 변환한다. 본 연구에서는 최신 딥러닝 기반 STT 모델(e.g., Whisper, Conformer, DeepSpeech 등)인 $SST(T_S) \rightarrow Transcript_s$을 활용할 수 있으며, 이들은 억양, 배경 소음, 화자 특성에 강건한 성능을 보여준다. 이렇게 얻어진 텍스트는 이후 장면의 의미 분석 및 검색 질의와의 매칭에 활용된다.

마지막 단계인 행동 구간과의 정합성 통합에서는 음성 정보만으로 장면 의미를 정확히 판단하기 어려운 경우, 시각적 행동 정보와의 통합이 필요하다. 이를 위해 본 연구는 발화 구간과 시각적 행동 구간 간의 시간적 중첩 여부를 기준으로 멀티모달 정보를 연계한다. 시간 구간이 다음과 같이 정의될 때,

- 행동 구간: $S_i = [s_i, e_i]$
- 발화 구간: $T_s = [t_{start}, t_{end}]$

겹치는 구간 $O_i = S_i \cap T_s \neq \emptyset$일 경우, 해당 구간에 대한 시각적 씬 그래프와 발화 텍스트를 매핑하며, 행동 구간과 발화 구간의 시간 매핑 예시에 따른 그래프 정보는 그림 2과 같고 상세 예시는 표 2와 같다.

표 2 행동 구간별 인식 결과 및 그래프 정보
Table 2 Recognition Results and Graph Information by Action Segment

행동 구간	시간 구간 [초]	인식된 행동	발화 구간 ID	발화 텍스트 예시	정합 여부
S_1	00:01 ~00:05	앉기	T_1	“여기 앉아도 돼요?”	O
S_2	00:06 ~00:09	책을 집어듦	T_2	“이 책, 예전에 읽어본 적 있어요.”	O
S_3	00:10 ~00:14	침묵+눈물 흘림	T_3	무발화(비어 있음)	X
S_4	00:15 ~00:19	고개를 끄덕임	T_4	“응, 나도 그렇게 생각해.”	O

O: 행동 구간과 발화 구간이 시간적으로 겹치고 의미적으로 연관됨
X: 해당 행동 구간에는 음성이 없거나 발화가 존재하지 않음

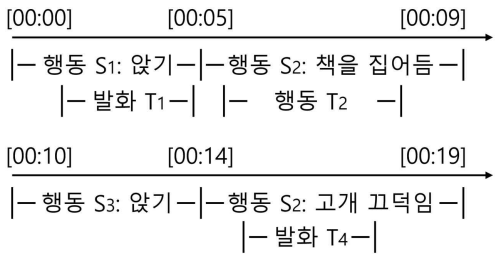

그림 2. 행동 구간별 인식 결과에 따른 그래프 정보
Fig. 2. Graph Information Based on Action-wise Recognition Result

3.4 행동 인식 기반 시각 정보 분석

앞선 단계에서 추출된 통합 씬 그래프(Scene Graph)와 발화 텍스트(STT 결과)는 시간 구간을 기준으로 정렬 및 병합되어, 장면 단위의 멀티모달 표현(Multimodal Representation)으로 통합된다.

멀티모달 표현은 다음과 같은 세 가지 요소를 포함한다.

1) 시각 정보: 행동 인식 기반의 통합 씬 그래프 – 객체와 관계 구조
2) 언어 정보: STT 기반의 발화 텍스트 – 문장 단위 또는 키워드 추출 결과
3) 시간 정보: 장면의 시작 및 종료 시간 – 질의 결과 재생 구간 지정

이러한 통합된 장면 표현은 검색 인덱스(database index)로 변환되어, 사용자의 다양한 질의 유형에 유연하게 대응할 수 있다. 제안하는 방법은 다음과 같은 유형의 질의를 지원한다.

1) 텍스트 기반 질의

사용자가 입력한 자연어 키워드를 발화 텍스트 또는 그래프의 객체/행동 태그와 매칭

예: “밥 먹자는 대사”, “엄마라는 단어가 나오는 장면”

2) 행동 중심 질의

시각적 씬 그래프 내 특정 관계 구조를 검색

예: “인물이 누군가를 안는 장면”, “눈물을 흘리는 장면”, “물건을 던지는 장면”

3) 복합 질의 (Multimodal Query)

시각+음성 정보를 함께 포함하는 의미 기반 질의

예: “혼잣말하면서 창밖을 바라보는 장면”, “웃으면서 ‘괜찮아’라고 말하는 장면”

이와 같은 인덱싱 구조를 통해 기존 키프레임 유사성 기반 검색이나 단순 메타데이터 검색 대비, 다음과 같은 고도화된 검색 성능을 기대할 수 있다

1) 맥락 유지: 동일 인물, 동일 행동이라도 발화 맥락에 따라 장면 분류 가능
2) 중복 제거: 동일 행동 반복 장면을 통합 표현함으로써 검색 효율성 향상
3) 사용자 중심 질의 대응: 직관적이고 문맥 중심의 질의 입력을 지원

결과적으로, 단순한 시각적 유사성 검색을 넘어 의미 기반 정밀 검색(semantic-level retrieval)이 가능한 멀티모달 검색 구조를 구현한다.

4. 멀티모달 장면 검색 기술의 응용 시나리오

4.1 개요

본 논문에서 제안하는 멀티모달 기반 장면 검색 기술은 OTT 콘텐츠의 다양한 소비 및 활용 환경에서 적용 가능하다. 특히 다음과 같은 실제 시나리오에서 정밀성과 활용성이 나타날 것으로 예상된다.

4.2 인물 중심 시점 전환 기능: 다른 인물 중심 다시보기

드라마나 예능 콘텐츠에서 특정 인물의 행동과 발화를 중심으로 다시보기를 원하는 사용자는 많지만, 기존 방법에서는 해당 인물의 전체 등장을 수작업으로 탐색해야 했다. 본 방법은 인물의 행동(예: 걷기, 웃기 등)과 이름 또는 대사 기반으로 장면을 필터링함으로써, “해당 인물이 말하거나 행동하는 구간만 연속 재생” 기능을 제공할 수 있다.

예시: “이도현이 혼잣말하는 장면만 모아서 보기”

4.3 감정 기반 장면 검색: 감정 중심 클립 탐색

행동 인식 및 STT 결과를 감정 분석과 연계하면, 감정 상태(예: 분노, 슬픔, 기쁨)에 해당하는 장면만 추출 가능하다. 이는 감정 중심 하이라이트 생성, 심리 분석 기반 시청, 혹은 특정 분위기 클립 편집 등에 유용하다.

예시: “주인공이 눈물을 흘리며 슬픈 대사를 하는 장면만 모아보기”

처리 방식: 행동 인식(‘우는 동작’) + STT(감정어 추출)

4.4 맥락 기반 발화 탐색: 특정 발화 검색

기존 자막 기반 검색은 발화 시간과 장면의 시각 정보가 분리되어 있어 실제 장면 탐색에 어려움이 있었다. 제안 방법은 발화와 시각 정보가 시간적으로 정합되도록 통합되어 있어, 사용자는 특정 대사나 문장을 입력함으로써 즉시 관련 장면을 탐색할 수 있다.

예시: “‘괜찮아, 넌 잘하고 있어’라는 말이 나오는 장면 찾아줘”

처리 방식: STT 인덱스에서 직접 텍스트 매칭 + 시각 씬 정보 제공

이러한 시나리오는 단순한 장면 탐색을 넘어 사용자 맞춤형 콘텐츠 소비, 감정/인물 기반 하이라이트 제작, 검색 기반 2차 저작물 생성 등의 실질적 응용을 가능하게 하며, 산업적 활용 가능성을 뒷받침할 것으로 기대된다.

5. 결론

본 논문에서는 OTT 콘텐츠 환경에서의 장면 탐색 정밀도를 향상시키기 위한 멀티모달 행동 및 음성 인식 기반의 장면 검색 고도화 방법을 제안하였다. 제안하는 방법은 Transformer 기반의 행동 인식 모델을 활용하여 영상의 연속된 프레임을 의미 있는 행동 구간으로 분할하고, 각 구간의 핵심 객체 및 관계를 추출하여 중복을 제거한 통합 씬 그래프를 구성함으로써 시각 정보의 의미적 밀도를 높였다. 또한, STT

(Speech-to-Text) 기술을 통해 인물의 발화 구간을 정밀하게 추출하고 이를 시각 정보와 시간적으로 정렬·매핑함으로써, 대사와 장면 간의 의미 정합성을 확보하였다.

이러한 멀티모달 통합 분석을 통해, 기존 방식보다 더 정밀하고 맥락 중심적인 장면 검색을 실현할 수 있으며, 이는 사용자의 검색 의도를 보다 정교하게 반영하는 기술적 진보로 평가된다. 또한 제안된 방식은 인물 중심 시점 전환, 감정 기반 클립 탐색, 특정 발화 기반 장면 검색 등 다양한 실용 시나리오에 적용 가능하며, OTT 플랫폼 내 개인화된 추천 방법, 시청자 중심의 인터랙티브 콘텐츠 탐색, 2차 저작물 제작 도구 등으로의 확장 가능성도 포함하고 있다.

향후 연구에서는 본 방법을 다양한 장르 및 형식의 영상 콘텐츠에 적용하여 범용성과 실효성을 검증하고, 자막, 얼굴 인식, 배경음 분석 등 추가적인 멀티모달 정보의 통합을 통해 보다 정교하고 직관적인 콘텐츠 검색 환경을 구축하는 방향으로 기술을 발전시켜 나갈 계획이다.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2025년도 문화기술 연구개발 사업으로 수행되었음(과제명 : OTT 콘텐츠 저작권 보호기술개발 및적용을 위한 저작권기술(+법) 융합인재양성, 과제번호 : RS-2023-00225267)

참 고 문 헌

[1] Sangwon Lee and Sunmi Lee, “The Impact of OTT Service Growth on the Market Performance of the Broadcasting Industry”, Journal of the Korea Contents Association, Vol. 22, No. 4, pp. 199 - 206, 2022. DOI: https://doi.org/10.5392/JKCA.2022.22.04

[2] Cheolmin Lim, Yoomi Jang, Heesoo Kim, Sunho Kim, Sanghyun Lee, and Seongcheol Kim, “A Study on Content Consumption Patterns through Domestic OTT Services: Focusing on Temporal and Spatial Contexts”, Journal of the Korea Digital Content Society, Vol. 24, No. 2, pp. 273 - 291, 2023. DOI: http://dx.doi.org/10.9728/dcs.2023.24.2.273

[3] Donghwan Noh, “Enhancing Competitiveness Through Data Utilization in Video OTT Platforms”, Media Issue & Trend, Korea Communications Agency (KCA), Vol. 51, pp. 30 - 33, 2022. https://www.kca.kr/Media_Issue_Trend/vol51/download/KCA_Media_Issue_Trend_vol51_featured_report_03.pdf

[4] Dohyung Park, “Segmentation and Data-Driven Persona Extraction through OTT Service Usage Logs”, Journal of the Korea Institute of Information Scientists and Engineers, Vol. 25, No. 3, pp. 35 - 45, 2023. DOI: https://doi.org/10.5762/KAIS.2024.25.3.312

[5] KBS Archive, “Introducing the Result of 1 Million Hours of KBS Archive”, Old TV: KBS Archive, 2025. https://www.youtube.com/watch?v=TsOA0PdrW2s

[6] Zhang Fuying, “Research on the Application of Scene Transition Technique in the Film The Myth”, Journal of the Korea Entertainment Industry Association, Vol. 11, No. 8, pp. 187 - 196, 2017. DOI: https://doi.org/10.21184/jkeia.2017.12.11.8.187

[7] Haowei Liu, Yaya Sh, Haiyang Xu, Chunfeng Yuan, Qinghao Ye, Chenliang Li, Ming Yan, Ji Zhang, Fei Huang, Bing Li, Weiming Hu, “Unifying Latent and Lexicon Representations for Effective Video-Text Retrieval”, Computer Vision and Pattern Recognition, 2024. DOI:

https://doi.org/10.48550/arXiv.2402.16769
[8] Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, Yihang Duan, Xinyu Lyu, Hengtao Shen, “Text-Video Retrieval with Global-Local Semantic Consistent Learning”, Computer Vision and Pattern Recognition, 2024. DOI: https://doi.org/10.48550/arXiv.2405.12710
[9] Electronics and Telecommunications Research Institute (ETRI), “Trends in AI-Based Video Content Generation Technologies”, ETRI Electronics and Telecommunications Trends Analysis, Vol. 34, No. 3, pp. 34 - 42, 2019. https://ettrends.etri.re.kr/ettrends/177/09051 77004/34-3_034-042.pdf
[10] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, Jun-Yan Zhu, “Generalizing Dataset Distillation via Deep Generative Prior”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. DOI: https://doi.org/10.1109/CVPR52729.2023.0036 4
[11] Manli Wang, Jiayue Li, Changsen Zhang, “Low-light image enhancement by deep learning network for improved illumination map”, Computer Vision and Image Understanding, Vol. 233, 2023. DOI: https://doi.org/10.1016/j.cviu.2023.103681
[12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision”, Audio and Speech Processing (eess.AS); Computation and Language (cs.CL); Machine Learning (cs.LG); Sound (cs.SD), 2022, DOI: https://doi.org/10.48550/arXiv.2212.04356
[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition”, Audio and Speech Processing (eess.AS); Machine Learning (cs.LG); Sound (cs.SD), 2020. DOI: https://doi.org/10.48550/arXiv.2005.08100
[14] Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, Khoa Luu, “HyperGLM: HyperGraph for Video Scene Graph Generation and Anticipation”, Computer Vision and Pattern Recognition (cs.CV), 2024. DOI: https://doi.org/10.48550/arXiv.2411.18042

저 자 소 개

박승현(Seung-Hyeon Park)

2023.2 한국방송통신대학교 컴퓨터과학과 학사
2024.3-현재 숭실대학교 컴퓨터학과 석사
<주관심분야> 저작권 보호 및 이용활성화

유인재(In-Jae Yoo)

2017.8 고려사이버대학교 소프트웨어공학과 학사
2022.2 숭실대학교 컴퓨터학과 석사
1923.2-현재 : 숭실대학교 컴퓨터학과 박사과정
<주관심분야> 저작권 보호 및 이용활성화

박병찬(Byeong-Chan Park)

2015.2 학점은행제 졸업
2018.2 숭실대학교 컴퓨터학과 석사
2023.8 숭실대학교 컴퓨터학과 박사
2023.9-현재 숭실대학교 초빙교수
<주관심분야> 저작권 보호 및 이용활성화

김석윤(Seok-Yoon Kim)

1980.2 서울대학교 전기전자 졸업
1990.2 University of Taxas at Austin Dept. of ECE 석사
1993.2 University of Taxas at Austin Dept. of ECE 박사
1982-1987 ETRI 연구원
1993-1995 모토로라 책임 연구원
1995-현재 : 숭실대학교 교수
<주관심분야> 저작권 보호 및 이용활성화

김영모(Young-Mo Kim)

2003.2 대전대학교 컴퓨터공학과 졸업
2005.2 대전대학교 컴퓨터공학과 석사
2011.2 대전대학교 컴퓨터공학과 박사
2012-현재 : 숭실대학교 교수
<주관심분야> 저작권 보호 및 이용활성화