Check for updates
특집논문 (Special Paper)
방송공학회논문지 제31권 제3호, 2026년 5월 (JBE Vol.31, No.3, May 2026)
https://doi.org/10.5909/JBE.2026.31.3.373
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

다중 객체 추적과 객체 재식별 기반 키프레임 추출 기술

이 승 복[a)], 민 병 석[a)‡]

An Enhanced Keyframe Extraction Technique based on Multi-Object Tracking and Re-Identification

SeungBok Lee[a)] and Byungseok Min[b)‡]

요 약

디지털 영상 데이터의 급증에 따라 효율적인 영상 관리 기술의 중요성이 증대되고 있다. 본 논문은 프레임별 상이한 객체 조합을 반영하면서 중복되지 않은 키프레임을 선별하기 위해 다중 객체 추적(MOT)과 객체 재식별(Re-ID)을 통합한 파이프라인을 제안한다. 특히 추적 기반 키프레임 추출의 핵심 한계인 Track ID 단절 문제를 해결하기 위해 슬라이딩 윈도우 버퍼 기반 Crop Histogram Track ID 복원, Hybrid Continuity Score 기반 그룹 경계 판단, dHash 기반 구조적 중복 제거를 적용하였다. 또한 RT-DETR/YOLO 기반 객체 탐지, BoTSORT 기반 추적, 바타차야 거리 기반 이중 히스토그램 필터링, OSNet 기반 재식별, k-조합 탐욕 알고리즘 기반 키프레임 선택, DPT 기반 단안 깊이 추정을 순차적으로 수행하였다. 실험 결과, 제안 방법은 기존 방식 대비 Track ID 연속성을 향상시키고 Re-ID 처리 시간을 단축하면서도 키프레임 대표성을 효과적으로 유지하였다. 또한 다양한 장면 변화와 객체 재등장 상황에서도 안정적인 키프레임 선별 성능을 확인하였다. 이는 대용량 영상 요약 및 관리 시스템에 효과적으로 활용될 수 있음을 보여준다. 실제 응용을 위한 활용 가능성도 확인하였다.

Abstract

With the rapid proliferation of digital video content, demand for efficient video management has grown substantially. This study proposes an integrated pipeline combining multi-object tracking (MOT) and re-identification (Re-ID) for effective keyframe selection representing essential video content. We identify Track ID discontinuity as a fundamental limitation of tracking-based keyframe extraction and address it through three complementary strategies: sliding window buffer-based Crop Histogram Track ID restoration, Hybrid Continuity Score-based group boundary detection, and dHash-based structural deduplication. The proposed system sequentially executes RT-DETR/YOLO-based object detection, BoTSORT-based multi-object tracking, enhanced frame grouping, Perceptual Hashing deduplication, Bhattacharyya distance-based dual histogram filtering, OSNet-based Re-ID, k-combination greedy keyframe selection, and DPT-based monocular depth estimation. Experimental results demonstrate that the proposed method achieves improved Track ID continuity and reduced Re-ID processing time while preserving keyframe representativeness relative to conventional pipelines.

Keyword : Keyframe Extraction, Object Detection, Object Tracking, Re-Identification, Depth Estimation

Ⅰ. 서 론

디지털 영상 콘텐츠의 폭발적 증가는 효율적인 영상 관리 및 분석 기술에 대한 수요를 급격히 증대시키고 있다. 특히 영상 검색, 비디오 요약, 3차원 장면(3D Scene) 재구성, 감시 시스템 등 다양한 분야에서 전체 프레임을 처리하는 것은 과도한 연산 비용과 중복 정보로 인해 비효율적이며, 영상의 핵심 내용을 대표하는 소수의 프레임을 효과적으로 선별하는 기술이 필수적이다[10,11,14,18,19]. 이러한 맥락에서 키프레임 추출(keyframe extraction)은 긴 비디오 시퀀스에서 영상의 핵심 내용을 대표하는 소수의 프레임을 선별하는 기술로, 효과적인 키프레임은 다양한 객체 조합과 장면 변화를 포함하면서도 시각적 중복은 최소화되어야 한다. 전통적인 키프레임 추출 방법론은 색상 히스토그램이나 텍스처 등 저수준(low-level) 특징에 의존하거나 단순한 군집화(clustering)에 그쳐[1], 객체의 의미적(semantic) 정보와 동적 변화를 반영하지 못하는 한계가 있었다. 최근 딥러닝 기반 객체 탐지(YOLOv26, RT-DETR), 다중 객체 추적(BoTSORT, ByteTrack), 객체 재식별(OSNet) 기술이 비약적으로 발전하고 있다[2,3,4,20,24]. 하지만 키프레임 추출이라는 통합 파이프라인 내에서 체계적으로 결합한 연구는 아직 부족한 실정이고, 특히 기존 추적 기반 접근법은 다음과 같은 근본적인 한계를 가진다.

첫째, 프레임 건너뛰기(frame_skip) 설정에 의한 추적 ID(Track ID) 단절이다. 처리 속도를 위해 일부 프레임을 건너뛸 경우, 트래커 입장에서 객체가 갑작스럽게 이동한 것으로 인식되어 새로운 ID를 할당하게 된다. 둘째, 프레임 건너뛰기를 제거하더라도 가림(occlusion), 빠른 움직임, 탐지 누락 등 트래커 자체의 한계로 ID가 재할당되는 문제가 근본적으로 존재한다. 셋째, 재식별 단계가 $O(N^2)$ 비교 연산으로 병목이 발생하여, 앞 단계에서 프레임이 충분히 줄어들지 않으면 전체 처리 시간이 급격히 증가하는 문제를 가진다. 본 연구에서는 이러한 문제들을 체계적으로 분석하고, 각 문제에 대응하는 다중 구성 요소를 통합한 파이프라인을 제안한다. 본 논문의 기술적 주요 기여 요소는 다음과 같이 제시된다.

① 다중 객체 추적(MOT)과 재식별(Re-ID)을 통합하는 키프레임 추출 파이프라인에서, Crop Histogram 기반 추적 ID(Track ID) 복원과 Hybrid Continuity Score를 도입하여 추적 ID 단절 문제를 해결한다. Crop Histogram은 탐지 경계 상자(detection bounding box)의 밝기/채도(brightness/saturation) 히스토그램을 비교하여, 객체 가림 후 위치가 크게 변한 경우에도 외형 유사성으로 동일 객체를 안정적으로 복원한다. 특히 최근 N개 프레임을 슬라이딩 윈도우 버퍼로 유지하고 현재 프레임과 순차적으로 매칭함으로써, 가림이 여러 프레임에 걸쳐 지속되는 경우에도 ID 복원이 가능하도록 한다.

② 객체 인스턴스 ID 기반 k-조합 탐욕 알고리즘으로 다양한 객체 상호작용 패턴을 포착하는 키프레임 선택 전략을 제안한다.

③ 바타차야(Bhattacharyya) 거리 기반 이중 히스토그램 필터링과 dHash(Difference Hash) 기반 Perceptual Hashing 중복 제거를 결합하여, 다단계 중복 제거 체계를 구축한다. dHash는 SSIM 대비 수십 배 빠른 속도로 구조적 중복을 제거하면서도 조명 변화에 강건한 장점을 가진다.

④ 선별된 키프레임에 대해 DPT 단안 깊이 추정을 적용하여 장면의 공간적 구조 정보를 함께 제공함으로써, 3D 재구성 등 다양한 후속 작업의 입력으로 활용할 수 있는 확장된 출력을 생성한다.

본 논문의 이후 구성은 다음과 같다. Ⅱ장에서 관련 연구를 검토하고, Ⅲ장에서 제안 방법을 상세히 기술한다. 이후 Ⅳ장에서 실험 결과를 분석하고, Ⅴ장에서 결론과 향후 연구 방향을 제시한다.

a) 세종대학교 인공지능학과(Sejong University)
b) 세종대학교 인공지능데이터사이언스학과(Sejong University)
‡ Corresponding Author : 민병석(Byungseok Min)
E-mail: bmin@sejong.ac.kr
Tel: +82-2-3408-3348
ORCID: https://orcid.org/0000-0001-9826-3471
※ 이 논문의 결과 중 일부는 한국방송·미디어공학회 2025년 동계학술대회에서 발표한 바 있음.
※ 이 논문은 2026년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2026-RISE-01-019-04)
· Manuscript March 11, 2026; Revised April 22, 2026; Accepted April 23, 2026.

Ⅱ. 관련 연구

1. 키프레임 추출 방법

키프레임 추출 연구는 초기에 인접 프레임 간 픽셀 차이나 색상 히스토그램 변화를 기준으로 샷 경계(shot boundary)를 탐지하는 방식에서 출발하였다. 이러한 저수준(low-level) 특징 기반 접근은 구현이 단순하다는 장점이 있으나, 영상 내 객체의 의미적 변화를 반영하지 못한다는 근본적 한계가 존재한다. 이를 보완하기 위하여 텍스처, 엣지, 구조적 유사도 등 다수의 시각 특징을 융합하고 군집화(clustering) 기법을 결합하는 방향으로 기술들이 제안되어 왔다. Kaur는 색상, 텍스처, 엣지, 구조적 유사도 네 가지 특징을 융합하고 Fuzzy-C Means 클러스터링을 결합함으로써 단일 특징 대비 견고한 성능을 보고하였다[1]. 그러나 K-means, DBSCAN 등 군집화 기반 방법은 프레임 간 시간적 순서 정보가 소실되고 클러스터 수 결정이 어려워 영상 특성에 따른 성능 편차가 크다는 문제가 잔존하였다. 광학 흐름(optical flow) 기반 방법 또한 카메라 자체의 움직임과 객체의 움직임을 구분하지 못하여 카메라 이동이 빈번한 영상에서 신뢰성이 저하되었다. Truong 등은 비디오 추상화 분야의 광범위한 문헌 고찰을 통해, 저수준 특징에만 의존하는 방법론으로는 영상 내 의미적 내용 변화를 충분히 표현하기 어려우며 객체 수준의 고수준(high-level) 정보를 활용하는 접근 필요성을 제안하였다[14].

2. 딥러닝 기반 객체 탐지

객체 탐지는 키프레임 추출 파이프라인의 첫 번째 단계로서, 각 프레임에서 객체의 위치, 클래스, 신뢰도 정보를 제공하는 핵심 모듈이다. 딥러닝 기반 탐지 연구는 초기에 R-CNN 계열의 2단계(two-stage) 검출 방법이 주도하였으며, Ren 등이 제안한 Faster R-CNN은 Region Proposal Network(RPN)를 도입하여 정확도를 크게 향상시켰다[7]. 그러나 2단계 검출 방법은 영역 제안과 분류를 순차적으로 수행하는 구조상 실시간 처리에 적합하지 않다는 한계가 지적되었다. 이를 극복하기 위하여 Redmon 등은 전체 이미지를 단일 신경망으로 처리하는 1단계(one-stage) 탐지기 YOLO를 제안하였으며, 이후 YOLOX 등 후속 연구를 거쳐 실시간 처리와 정확도의 균형을 지속적으로 개선하였다[8,12]. 가장 최근 모델인 YOLOv26은 분포 초점 손실(Distribution Focal Loss, DFL) 제거와 종단간(end-to-end) NMS 비사용 추론 구조를 도입하고, 소형 객체 인식 레이블 할당(STAL) 및 MuSGD 옵티마이저를 채택하여 엣지 및 저전력 장치에서의 탐지 효율과 정확도를 동시에 향상시켰다[2]. 한편 Zhao 등이 제안한 RT-DETR은 Transformer 기반 탐지기로서 DETR 계열의 느린 수렴 문제를 해결하고 YOLO 수준의 처리 속도에서 높은 탐지 정확도를 달성하였다[24]. 본 연구에서는 탐지 성능과 처리 속도에 대한 부분을 고려하여 YOLOv26을 기본 탐지 모델로 파이프라인을 구성하였다.

3. 다중 객체 추적(Multi-Object Tracking)

다중 객체 추적(MOT)은 연속된 프레임에서 동일 객체에 일관된 추적 ID(Track ID)를 부여하는 기술로, 키프레임 추출 파이프라인에서 프레임 간 객체 구성 변화를 추적하는 핵심 기반을 제공한다. 다중 객체 추적(MOT) 연구는 초기에 위치 정보만을 활용하는 방식에서 출발하였다. Bewley 등이 제안한 SORT는 칼만 필터(Kalman Filter)로 객체 위치를 예측하고 헝가리안(Hungarian) 알고리즘으로 탐지 결과를 연결하는 간결한 구조로 빠른 처리 속도를 달성하였으나, 외관 정보를 전혀 활용하지 않아 가림(occlusion)이나 교차 상황에서 ID 스위칭이 빈번히 발생하는 한계가 있었다[6]. 이를 보완하기 위하여 Wojke 등은 DeepSORT를 제안하여 재식별(Re-ID) 기반 외관 특징 벡터를 추적에 통합함으로써 ID 스위칭 문제를 상당 부분 완화하였다[5]. 이후 Zhang 등은 ByteTrack을 제안하여, 높은 신뢰도 탐지 결과뿐 아니라 낮은 신뢰도 탐지 결과까지 2단계로 활용하는 BYTE 연관(association) 전략을 도입함으로써 가림(occlusion) 상황에서 미매칭 트랙의 복구 능력을 크게 향상시켰다[4]. 또한 Aharon 등은 BoTSORT를 제안하여, ByteTrack에 카메라 모션 보상(camera motion compensation, CMC)을 칼만 필터(Kalman Filter) 예측 단계에 통합하고 재식별(Re-ID) 외관 특징을 결합함으로써 카메라 이동 환경에서도 강건한 추적 성능을 달성하였다[20]. 그러나 BoTSORT를

포함한 현재의 추적기들은 가림(occlusion) 이후 탐지가 재개될 때 동일 객체에 새로운 ID를 할당하는 ID 재할당 문제를 근본적으로 해결하지 못하고 있다. 본 연구에서는 BoTSORT를 기본 추적기로 채택하고, 이러한 ID 재할당 문제를 후처리 단계에서 Crop Histogram 기반 복원을 통해 보완하는 방식을 채택하였다.

4. 객체 재식별(Re-Identification)

객체 재식별(Re-ID)은 서로 다른 시점이나 촬영 환경에서 취득된 동일 객체를 인식하는 기술로, 다중 객체 추적에서 발생하는 ID 단절 문제를 보완하는 수단으로 활발히 연구되어 왔다. 초기 재식별(Re-ID) 연구는 색상 히스토그램, LBP(Local Binary Pattern) 등 수작업으로 설계된 특징(hand-crafted feature)에 의존하였으나, 조명 변화와 자세 변동에 취약한 한계가 있었다. 이후 딥러닝 기반 metric learning 방법이 도입되어, 삼중 손실(triplet loss) 및 대조 손실(contrastive loss)을 활용한 임베딩 학습이 재식별(Re-ID) 성능을 크게 향상시켰다. 이 과정에서 Zheng 등의 Market-1501 등 대규모 공개 벤치마크가 구축되어 공정한 성능 비교가 가능해졌으며, Luo 등은 강력한 학습 기준선(baseline)과 배치 정규화 기법으로 성능을 추가로 개선하였다[13,9]. 이러한 발전을 바탕으로 Zhou 등은 다중 스케일 특징을 통합적으로 학습하는 경량 구조의 OSNet(Omni-Scale Network)을 제안하여, 512차원 임베딩 벡터를 기반으로 코사인 유사도 매칭을 수행하고 Market-1501 등의 벤치마크에서 높은 성능을 달성하였고, 본 연구에서도 OSNet 모델을 기반으로 객체 재식별에 사용하였다[3].

5. 시각적 중복 제거(Visual Redundancy Elimination)

이미지 중복 검출은 초기에 픽셀 단위 비교(MSE) 또는 SSIM(Structural Similarity Index) 기반 방법이 주로 활용되었다. 이러한 방법은 높은 정밀도를 제공하지만 $O(W\times H)$의 연산 비용이 요구되어 대용량 프레임 집합에 대한 실시간 처리에는 적합하지 않다는 한계가 있었다. 이를 해결하기 위하여 Perceptual Hashing 기법이 제안되었으며, 이미지의 구조적 특징을 짧은 고정 길이 비트열로 인코딩함으로써 해시 간 해밍 거리(Hamming distance) 비교만으로 O(1) 시간 복잡도의 유사도 판별을 가능하게 하였다. 초기에 제안된 aHash(Average Hash)는 이미지를 축소한 후 전체 픽셀의 평균값을 기준으로 0/1 비트를 결정하는 단순한 방식으로, 구현이 용이하지만 전체적인 밝기 변화에 민감하여 조명 변동이 있는 환경에서 신뢰성이 저하되는 문제가 있었다. 이를 보완하기 위하여 dHash(Difference Hash)가 제안되었으며, 인접 픽셀 간의 상대적 밝기 차이(좌→우)를 비트로 인코딩함으로써 전체적인 조명 변화에 강건하고 이미지의 구조적 패턴을 효과적으로 포착하는 것으로 알려져 있다. 64bit 고정 길이 해시를 사용하므로 SSIM 대비 수십~수백 배 빠른 연산 속도를 제공하며, 본 연구에서는 이러한 장점을 바탕으로 dHash를 채택하여 Enhanced Primary Selection 출력 프레임 집합 내의 구조적 중복을 효율적으로 제거하였다.

6. 단안 깊이 추정

단안 깊이 추정(monocular depth estimation)은 단일 2D 영상으로부터 각 픽셀의 깊이 값을 예측하는 기술로, 키프레임의 공간적 구조 정보를 함께 제공함으로써 후속 3차원 재구성 등 후속 응용(downstream task)의 입력 품질을 향상시키기 위해 활용된다[11,18,19]. 초기에는 LiDAR나 스테레오 카메라 등 별도의 하드웨어 장치에 의존하는 방식이 주를 이루었으나, 범용적 적용에 한계가 있었다. 이후 CNN 기반 단안 깊이 추정 방법이 제안되어 단일 이미지만으로 깊이 예측이 가능해졌으며, Ranftl 등은 DPT(Dense Prediction Transformer)를 제안하여 Vision Transformer를 밀집 예측(dense prediction)에 적용함으로써 세밀하고 일관된 깊이 맵 생성 성능을 향상시켰다[15]. 또한 Ranftl 등은 MiDaS를 통해 다양한 데이터셋을 혼합 학습하는 방식으로 특정 도메인에 의존하지 않는 범용적인 깊이 추정 성능을 달성하였다[16]. 최근에는 Yang 등이 Depth Anything을 제안하여 대규모 비지도 사전학습을 통해 다양한 환경에서 범용적인 깊이 추정 성능을 달성하고 있다[17]. 본 연구에서는 선별된 키프레임에 DPT를 적용하여 깊이 맵을 추출하고, Otsu 임계화를 통해 객체 바운딩 박스 영역 내 전경과 배경을 자동

으로 분리한 뒤 전경 픽셀의 깊이 통계를 산출함으로써, 2D 프레임과 함께 각 객체의 공간적 깊이 정보를 제공하는 확장된 출력을 생성하였다.

Ⅲ. 제안 방법

1. 시스템 개요

제안하는 다중 객체 추적 기반 키프레임 추출 파이프라인의 전체 구조는 그림 1에 나타나 있다. 파이프라인은 크게 비디오 프레임 입력, 객체 탐지 및 추적, 중복 제거 및 재식별, 그리고 최종 키프레임 선택 및 객체별 3D 깊이(Depth) 정보 추출의 단계로 구성된다. 이를 위한 구체적인 단계별 수행 절차 및 수식적 전개는 알고리즘 1에 상세히 기술되어 있다.

제안 시스템은 그림 1과 알고리즘 1에서와 같이, RT-DETR 또는 YOLOv26과 같은 탐지 모델과 OSNet, DPT 모델을 로딩하여 초기화한다. 이후 입력된 비디오 시퀀스에 대하여 'Enhanced Primary Selection'을 수행한다. 이 단계에서는 모든 프레임을 대상으로 객체를 탐지 및 추적하고, Crop Histogram 기반 추적 ID(Track ID) 복원으로 단절된 ID를 재연결한다. 동시에 Hybrid Continuity Score를 계산하여 프레임 그룹의 경계를 판단하고, 각 그룹 내에서 탐지 품질이 가장 높은 프레임을 대표로 선별한다. 이어서 선별된 대표 프레임들에 대해 다단계 중복 제거를 수행한다. 먼저 Perceptual Hashing(dHash)을 적용하여 구조적으로 유사한 프레임을 빠르게 걸러내고, 바타차야(Bhattacharyya) 거리 기반의 이중 히스토그램 필터링(Profile Tracking)을 통해 시각적 중복을 추가로 제거한다.

이후 OSNet 모델을 활용하여 'Duplicate Scene Grouping

알고리즘 1. 다중 객체 추적 기반 키프레임 추출 파이프라인
Algorithm 1. Multi-Object Tracking-Based Keyframe Extraction Pipeline

Algorithm 1 Multi-Object Tracking-Based Keyframe Extraction Pipeline

Require: Video $V = \{f_1, \ldots, f_T\}$, models $\mathcal{M}_{\text{det}}, \mathcal{M}_{\text{reid}}, \mathcal{M}_{\text{depth}}$, thresholds $\tau_{\text{crop}}, \tau_{\text{hybrid}}, \tau_{\text{hash}}, \tau_{\text{hist}}, \tau_{\text{post}}, \tau_{\text{reid}}, \tau_{\text{merge}}, N_{\max}, N_{\text{buf}}$
Ensure: Keyframes K with per-object depth statistics Ψ

/* **Step 1: Enhanced Primary Selection** */
$G \leftarrow \emptyset; \quad F \leftarrow \emptyset$
for each $f_t \in V$ **do**
  $D_t \leftarrow \text{Detect}(f_t; \mathcal{M}_{\text{det}}); \quad T_t \leftarrow \text{BoTSORT}(D_t)$
  $T_t \leftarrow \text{CropHistRestore}(T_t, \text{Buffer}_{t-N_{\text{buf}}:t-1}, \tau_{\text{crop}})$ ▷ Eq.(1),(2)
  $S_{\text{hybrid}} \leftarrow w_J \frac{|T_{t-1} \cap T_t|}{|T_{t-1} \cup T_t|} + w_H \exp(-D_B(h_{t-1}, h_t)) + w_{\text{IoU}} \overline{\text{IoU}}$ ▷ Eq.(3)
  if $S_{\text{hybrid}} < \tau_{\text{hybrid}}$ **then**
    $F \leftarrow F \cup \{\arg\max_{f \in G}(|D_f| + \varepsilon \cdot \frac{1}{|D_f|} \sum s_i)\}$ ▷ Eq.(4)
    $G \leftarrow \{f_t\}$
  else
    $G \leftarrow G \cup \{f_t\}$
  end if
end for
if $G \neq \emptyset$ **then** $F \leftarrow F \cup \{\arg\max_{f \in G}(|D_f| + \varepsilon \cdot \frac{1}{|D_f|} \sum s_i)\}$
end if
/* **Step 2 & 3: Hashing & Histogram Deduplication** */
$F \leftarrow F \setminus \{f_{i+1} \mid D_H(b_i, b_{i+1}) \leq \tau_{\text{hash}} \wedge |D_{i+1}| \leq |D_i|\}$ ▷ dHash, Eq.(6)
$F \leftarrow F \setminus \{f_j \mid D_B(h_i, h_j) < \tau_{\text{hist}} \wedge (T_i \subseteq T_j \vee T_j \subseteq T_i)\}$ ▷ Profile, Eq.(8)
/* **Step 4: Re-Identification & Semantic Deduplication** */
$e_i \leftarrow \text{Embed}(\text{crop}_i; \mathcal{M}_{\text{reid}})$ for all $\text{crop}_i \in f \in F$
Assign Global IDs by greedy clustering $\{e_i\}$ per class with τ_{reid} ▷ Eq.(9)
Let O_t be the unified object set in f_t. For any pair $(f_i, f_j) \in F$ where $i < j$:
if $|D_i| = |D_j| \wedge |\text{common}| > 0 \wedge \frac{1}{|D_i|} \sum_{o \in \text{common}} \mathbf{1}[S_C(e_o^i, e_o^j) > \tau_{\text{merge}}] > 0.95$ **then**
  $F \leftarrow F \setminus \{f_j\}$
end if
/* **Step 5: Greedy Keyframe Selection** */
$C_t \leftarrow \{(o_u, o_v) \mid o_u, o_v \in O_t, u < v\}$ for each $f_t \in F$ ▷ 2-combinations, Eq.(10)
$S \leftarrow \emptyset; \quad K \leftarrow \emptyset$
while $|K| < N_{\max}$ **and** $\exists f_t \in F$ s.t. $C_t \setminus S \neq \emptyset$ **do**
  $f^* \leftarrow \arg\max_{f \in F} |C_f \setminus S|$
  $K \leftarrow K \cup \{f^*\}; \quad S \leftarrow S \cup C_{f^*}$
end while
/* **Step 6: Post-Processing & Depth Extraction** */
$K \leftarrow \text{PostProfileFilter}(K, \tau_{\text{post}})$
for each $f \in K$ **do**
  depth_map $\leftarrow \mathcal{M}_{\text{depth}}(f)$
  for each object bounding box $b_i \in D_f$ **do**
    Apply OtsuThreshold to normalize(depth_map[b_i])
    $\Psi_i \leftarrow \{\mu, \sigma, p_{25}, p_{50}, p_{75}\}$ of foreground depth pixels
  end for
end for
return K, Ψ

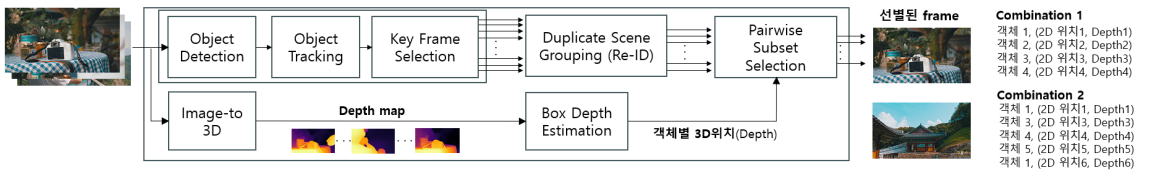

그림 1. 제안하는 다중 객체 추적 기반 키프레임 추출 파이프라인의 전체 구조
Fig. 1. Overall architecture of the proposed multi-object tracking-based keyframe extraction pipeline

(Re-ID)' 단계에 해당하는 객체 재식별을 수행하는데, 서로 다른 Track ID를 가진 동일 객체를 하나의 전역 ID(Global ID)로 통합하며, 의미론적 중복 프레임을 배제한다. 다음으로 'Pairwise Subset Selection' 단계에서는 k-조합 탐욕 알고리즘을 적용하여, 영상 내 등장하는 다양한 객체 조합을 최대한 포괄할 수 있는 핵심 키프레임을 최종적으로 선택한다.

마지막으로 2차 히스토그램 필터링을 거친 최종 키프레임 집합에 대하여, DPT 기반 단안 깊이 추정 모듈(Image-to-3D)을 적용한다. 이를 통해 각 키프레임 내 객체들의 3D 위치(Depth) 정보를 추출(Box Depth Estimation)함으로써, 알고리즘 1의 최종 출력물인 키프레임 집합과 객체별 깊이 통계 정보를 도출하며 파이프라인을 마무리한다.

2. 객체 탐지 및 다중 객체 추적

각 프레임 f_t에 RT-DETR 또는 YOLOv26을 적용하여 객체 집합 $D_t = \{d_1, d_2, ..., d_n\}$을 탐지한다. 각 탐지 d_i는 바운딩 박스 $b_i = (x1, y1, x2, y2)$, 클래스 레이블 c_i, 신뢰도 s_i를 포함한다. 탐지 결과는 BoTSORT 추적기에 입력되어 프레임 간 일관된 추적 ID(Track ID)가 할당된다. 추적 안정성을 높이기 위해 모든 프레임을 프레임 스킵 없이 탐지 및 추적에 활용한다. 프레임을 건너뛸 경우 트래커 입장에서 객체가 여러 프레임 분의 거리를 갑작스럽게 이동한 것으로 인식되어 IoU 매칭이 실패하고 새로운 ID가 할당된다. 전 프레임 처리는 탐지 횟수를 늘리지만, ID 연속성 개선으로 불필요한 그룹 생성이 줄어들어 후속 단계에 전달되는 대표 프레임 수가 감소하므로 전체적인 처리 효율을 유지하도록 하였다.

2.1 Box profile 기반 추적 ID(Track ID) 유지

모든 프레임을 처리하더라도 가림(occlusion), 빠른 움직임, 탐지 누락, ID 교환(ID swap) 등의 상황에서 트래커가 동일 객체에 새로운 ID를 부여하는 문제는 근본적으로 발생한다. 특히, 단일 직전 프레임과의 비교만으로는 가림 현상이 2프레임 이상 지속되는 경우 추적 ID 복원이 불가능한 한계가 존재한다. 이를 해결하기 위해, 최근 일정 개수의 프레임을 슬라이딩 윈도우 버퍼로 유지하고, 버퍼 내 각 프레임의 탐지 경계 상자(Detection Bounding Box) 영역에 대한 밝기(Brightness) 및 채도(Saturation) 결합 히스토그램을 현재 프레임과 비교하는 방식을 제안한다. 제안하는 크롭(Crop) 히스토그램 산출 방식은 식 (1)과 같다.

식 (1)에서 크롭 영역의 결합 히스토그램은 밝기 히스토그램과 채도 히스토그램의 가중합으로 정의되며, 이때 사용되는 가중치는 0.5를 기본값으로 설정한다. 이렇게 산출된 히스토그램을 바탕으로 단절된 추적 ID를 매칭하는 과정은 식 (2)로 정의된다.

식 (2)는 현재 프레임에 존재하는 특정 객체의 추적 ID를 복원하기 위해, 버퍼에 저장된 이전 과거 프레임들의 객체들 중 클래스가 동일하고 바타차야(Bhattacharyya) 거리가 사전에 정의된 허용 임계값보다 작은 후보군을 찾는 과정을 나타낸다. 복수의 매칭 후보가 존재할 경우, 히스토그램 거리가 가장 짧은 쌍부터 탐욕(Greedy) 방식으로 1:1 매칭을 수행하며, 시간적으로 가장 가까운 프레임부터 순차 탐색하여 오매칭 위험을 최소화한다. 이러한 일련의 추적 ID 복원 과정은 그림 2에 상세히 나타나 있다. 그림 2에서 볼 수 있듯이, Frame 40에서 ID_1로 추적되던 객체가 Frame 65에서 기둥에 의해 완전히 가려진(Occlusion) 후 Frame 80에서 다시 등장하면, 기존 트래커는 이를 새로운 객체인 ID_51로 오인하게 된다. 하지만 제안하는 방법을 통해 Frame 40과 Frame 80의 객체 영역 Histogram Profiling을 비교하여 두 객체의 외형적 유사성을 확인하고, 최종적으로 Frame 80의 객체 ID를 ID_1로 복원(Tracking ID Update)함으로써 추적의 연속성을 강건하게 유지할 수 있다.

$$h_{\text{crop}} = \alpha \cdot h_{\text{bright}} + (1-\alpha) \cdot h_{\text{sat}}, \quad \text{where } \alpha \in [0,1] \text{ (default } \alpha = 0.5) \tag{1}$$

$$\varphi(i) = \arg\min_j \; D_B\left(h^{(t)}_{\text{crop},i}, h^{(t')}_{\text{crop},j}\right) \text{ s.t. } D_B < \tau_{\text{crop}},\; c_i = c_j,\; t' \in \{t - N_{\text{buf}}, \ldots, t-1\} \tag{2}$$

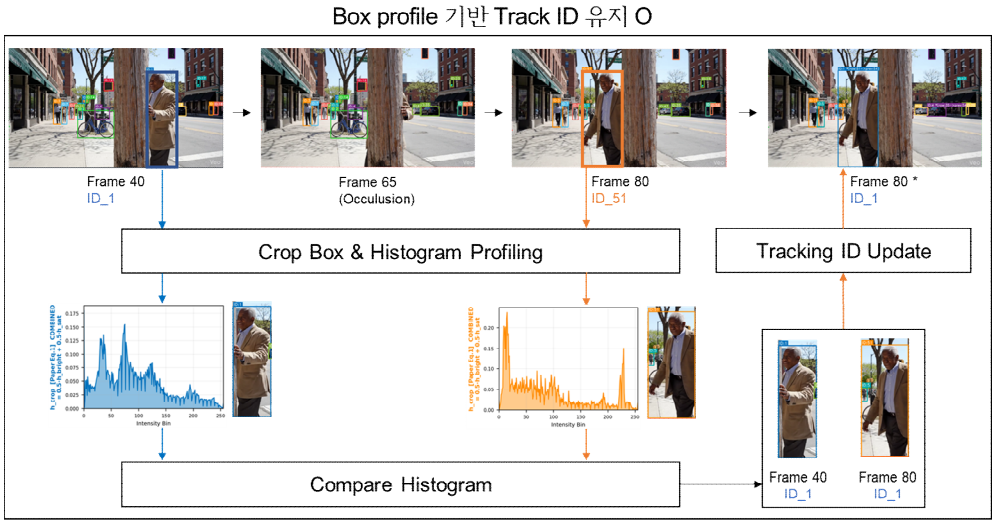

그림 2. 슬라이딩 윈도우 버퍼 기반 Crop Histogram Track ID 복원 과정
Fig. 2. Sliding window buffer-based Track ID restoration via crop histogram matching

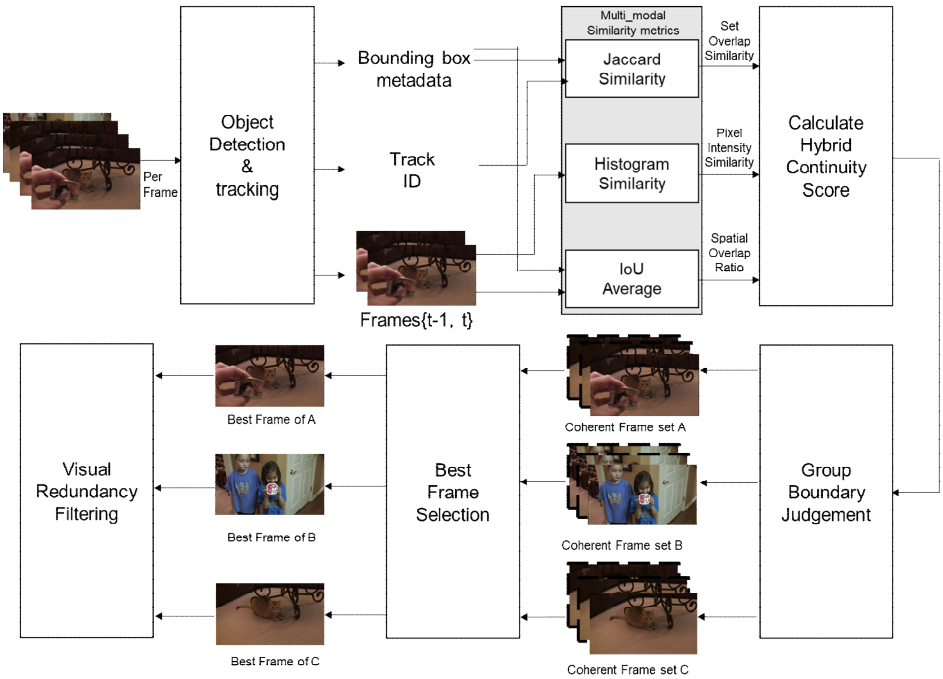

그림 3. Hybrid Continuity Score 기반 프레임 그룹 경계 판단 및 대표 프레임 선택 구조
Fig. 3. Frame group boundary detection and best-quality frame selection based on Hybrid Continuity Score

3. Enhanced Primary Selection

객체 탐지 및 추적 단계를 거친 후, 비디오의 중복성을 줄이고 핵심 프레임을 선별하기 위해 Enhanced Primary Selection 과정을 수행한다. 이 과정은 그림 3에 나타난 바와 같이, 하이브리드 연속성 점수(Hybrid Continuity Score)

를 기반으로 프레임 그룹의 경계를 판단하고 각 그룹에서 품질이 가장 높은 최적의 프레임을 선별하는 단계로 구성된다.

3.1 Hybrid Continuity Score

트래킹 ID 집합의 완전 일치 여부만으로는 프레임 그룹의 경계를 정확히 판단하기 어렵다. 이를 해결하기 위해 본 연구에서는 세 가지 보완적 지표를 결합한 하이브리드 연속성 점수를 제안하여 프레임 간 연속성을 판단한다. 제안하는 하이브리드 연속성 점수 산출 방식은 식 (3)과 같다.

$$S_{\text{hybrid}} = w_J \cdot \frac{|T_{t-1} \cap T_t|}{|T_{t-1} \cup T_t|} + w_H \cdot \exp(-D_B(h_{t-1},\, h_t)) + w_{\text{IoU}} \cdot \overline{\text{IoU}} \tag{3}$$

식 (3)에서 첫 번째 항은 인접 프레임 간 추적 ID 집합의 자카드 유사도(Jaccard similarity)를 의미한다. 두 번째 항은 프레임 전체 히스토그램 벡터 간의 바타차야(Bhattacharyya) 거리를 지수 함수로 변환한 시각적 유사도이며, 세 번째 항은 동일 클래스 탐지 객체 간의 평균 최적 IoU를 나타낸다. 각 지표에 곱해지는 가중치들은 이 세 가지 특성이 상호 보완적으로 작용하도록 조율한다. 가림 현상 등으로 추적 ID가 손실되어 자카드 유사도(Jaccard similarity)가 낮게 측정되더라도, 동일한 장면일 경우 히스토그램 유사도와 IoU가 높게 나타나 전체 하이브리드 점수는 사전에 설정된 임계값 이상을 유지하며 같은 그룹으로 묶이게 된다. 반면, 실제 카메라 시점 변경이나 장면 전환이 발생한 경우에는 세 지표가 모두 낮아져 새로운 그룹으로 명확히 분리된다.

3.2 최적 프레임(Best frame) 선택

프레임 그룹이 분리된 후, 각 그룹을 대표할 최적의 프레임을 선별하는 과정이 필요하다. 기존의 단순 중앙 프레임 선택 방식에서 벗어나, 본 연구에서는 각 그룹 내에서 탐지(Detection) 품질이 가장 높은 고품질 프레임을 대표로 선택한다. 이를 위해 프레임 내 탐지된 객체의 수를 1순위 기준으로 삼고, 탐지 수가 동일할 경우 평균 탐지 신뢰도(Confidence) 점수를 타이브레이커(Tie-breaker)로 적용하는 단일 최적화 수식을 설계하였으며, 이는 식 (4)와 같다.

$$f^* = \arg\max_{f_t \in G} \left(|D_t| + \varepsilon \cdot \frac{1}{|D_t|} \sum_{i=1}^{|D_t|} s_i \right) \tag{4}$$

식 (4)의 엡실론 항은 충분히 작은 양수 상수로서 동점 처리 계수 역할을 한다. 이 수식을 통해 탐지된 객체 수가 동일한 그룹 내 프레임들 사이에서 평균 탐지 신뢰도 점수가 선택에 반영되도록 유도함으로써, 모션 블러(Motion blur)나 탐지 누락이 가장 적고 객체가 가장 선명하게 포착된 최적의 단일 프레임이 최종 선별된다.

4. Perceptual Hashing 기반 중복 제거

Enhanced Primary Selection 단계에서 선택된 대표 프레임들 사이의 구조적 중복을 빠르게 걸러내기 위해, 본 연구에서는 dHash(Difference Hash) 기반의 지각적 해싱(Perceptual Hashing) 기법을 적용한다. dHash 연산을 위해 먼저 원본 이미지를 가로 N+1, 세로 N의 크기로 축소한 후 회색조(Grayscale) 이미지로 변환한다. 이후 식 (5)와 같이 인접한 픽셀 간의 밝기 차이를 연속적으로 비교하여 이진화된 비트(Bit) 값을 생성한다.

$$b_{x,y} = \mathbf{1}[I(x,\, y+1) > I(x,\, y)],\quad x \in \{0, \ldots, N-1\},\quad y \in \{0, \ldots, N-1\} \tag{5}$$

식 (5)의 $\boldsymbol{I}$ (x, y)는 축소된 회색조 이미지의 특정 좌표에 위치한 픽셀의 밝기 값을 의미한다. 가로 방향으로 인접한 두 픽셀을 비교하여 오른쪽 픽셀의 밝기가 왼쪽 픽셀보다 클 경우 1을, 그렇지 않을 경우 0을 할당한다. 본 연구에서는 기준 해시 크기 N을 8로 설정하였으며, 이에 따라 9x8 해상도로 축소된 이미지로부터 프레임당 총 64비트의 이진 해시 벡터가 생성된다. 이렇게 생성된 두 프레임 간의 해시 벡터를 비교하여 구조적 유사도를 판별하는 과정은 식 (6)과 같다.

$$D_H(b_1,\, b_2) = \sum_i \mathbf{1}[b_{1,i} \neq b_{2,i}] \tag{6}$$

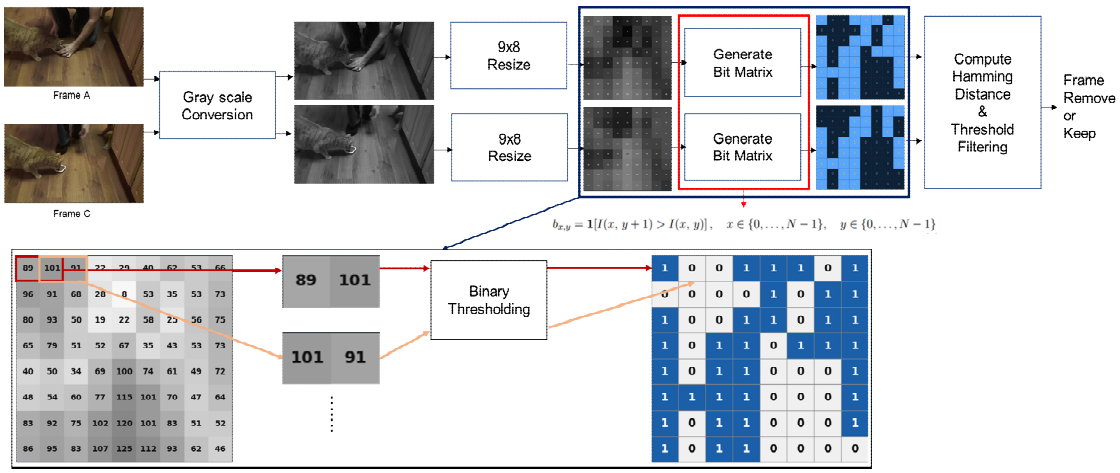

그림 4. dHash 기반 Perceptual Hashing을 통한 구조적 중복 프레임 제거 과정
Fig. 4. Structural deduplication process via dHash-based perceptual hashing

식 (6)은 두 프레임의 해시 벡터 사이에서 서로 값이 일치하지 않는 비트의 총 개수를 합산하여 해밍 거리(Hamming distance)를 산출하는 수식이다. 계산된 해밍 거리가 사전에 정의된 허용 임계값(본 연구에서는 10으로 설정) 이하일 경우, 두 프레임은 시각적 구조가 매우 유사한 중복 장면으로 판정된다. 중복으로 판정된 프레임 쌍에 대해서는 프레임 내에 탐지된 객체의 수가 더 많은 쪽을 정보량이 풍부한 프레임으로 간주하여 우선적으로 유지하고, 상대적으로 객체 수가 적은 프레임을 최종 제거한다. 이러한 일련의 dHash 기반 구조적 중복 제거 알고리즘의 세부 동작 과정은 그림 4에 상세히 나타나 있다. 그림 4에서 볼 수 있듯이, 9x8 해상도로 극소형화된 이미지에서 추출된 비트 행렬을 비교하는 방식은 영상의 전반적인 조명 변화나 국소적 노이즈에 강건하게 대응하면서도 프레임의 핵심적인 공간적 구조 패턴을 효과적으로 포착한다. 이를 통해 연산량이 큰 후속 재식별(Re-ID) 단계로 데이터가 전달되기 전에, 불필요한 시각적 중복 프레임들을 초고속으로 걸러내는 핵심 역할을 수행한다.

5. 히스토그램 기반 프레임 필터링(Profile Tracking)

Perceptual Hashing을 통한 구조적 중복 제거 이후에도 미세한 시각적 유사성을 띠는 중복 프레임들이 잔존할 수 있다. 이를 효과적으로 필터링하기 위해, 본 연구에서는 프레임의 밝기(Brightness)와 채도(Saturation) 정보를 결합한 히스토그램 프로파일링(Profile Tracking) 기법을 적용한다. 제안하는 결합 히스토그램 산출 수식은 식 (7)과 같다.

$$h_t = w_1 \cdot h_{\text{bright}}(f_t) + w_2 \cdot h_{\text{sat}}(f_t) \tag{7}$$

식 (7)에서 특정 프레임의 결합 히스토그램은 해당 프레임의 밝기 히스토그램과 채도 히스토그램에 각각의 가중치를 곱하여 합산한 결과로 정의된다. 두 가중치의 합은 1이 되도록 설정하며, 본 연구에서는 두 특성을 동등한 비율로 결합하였다. 이렇게 산출된 결합 히스토그램을 바탕으로 두 프레임 간의 시각적 분포 유사도를 정량적으로 측정하기 위해 식 (8)과 같이 바타차야(Bhattacharyya) 거리를 산출한다.

$$D_B(h_i, h_j) = -\ln\left(\sum_k \sqrt{h_i^{(k)} \cdot h_j^{(k)}}\right) \tag{8}$$

식 (8)은 두 히스토그램 벡터 간의 바타차야 거리를 계산하는 수식으로, 계산된 거리가 짧을수록 두 프레임의 시각적 분포가 매우 유사함을 의미한다. 파이프라인은 사전에 정의된 크기의 양방향 슬라이딩 윈도우 내에서 인접한 프레임 쌍들을 탐색하며, 산출된 바타차야 거리가 허용 임계

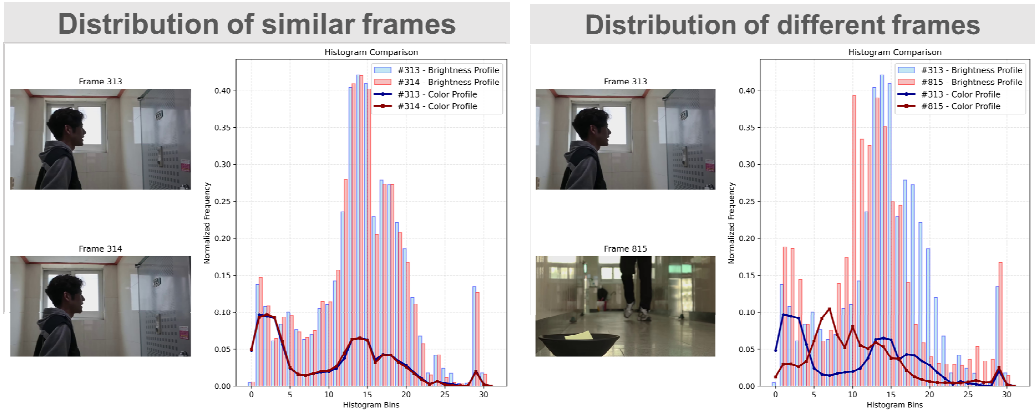

그림 5. Profile Tracking에서의 유사 프레임과 비유사 프레임 간 히스토그램 비교
Fig. 5. Histogram comparison between similar and dissimilar frames in profile tracking

값보다 작은 유사 프레임 쌍을 식별한다. 유사 프레임 쌍이 발견될 경우, 두 프레임에 포함된 추적 ID 집합 간의 부분집합(포함) 관계를 검사하여 상대적으로 적은 객체 정보를 담고 있는 프레임을 최종적으로 제거한다. 이러한 히스토그램 기반의 시각적 필터링 원리와 그 효과는 그림 5에 상세히 나타나 있다. 그림 5의 왼쪽 그래프에서 볼 수 있듯이, 시간적으로 인접하여 시각적으로 거의 동일한 두 프레임은 밝기 및 채도 히스토그램 분포가 거의 완벽하게 겹쳐지며 매우 짧은 바타차야 거리를 산출한다. 반면, 오른쪽 그래프와 같이 장면이 전환되거나 객체의 큰 움직임이 발생하여 시각적 차이가 명확한 두 프레임은 히스토그램의 형태가 확연히 어긋나게 분포함을 확인할 수 있다. 시스템은 이러한 정량적 분포 차이를 바탕으로 윈도우 내에 존재하는 불필요한 시각적 중복 프레임을 효과적으로 걸러낸다.

6. 객체 재식별(Re-Identification)

객체 재식별을 수행하기 위해 본 연구에서는 OSNet 모델을 활용한다. 선별된 각 키프레임 내에서 탐지된 개별 객체 영역(Detection crop)으로부터 512차원의 임베딩 벡터를 추출하며, 동일한 클래스(Class)에 속하는 객체 쌍에 대하여 코사인 유사도(Cosine similarity)를 계산한다. 이 과정은 식 (9)와 같이 정의된다[3].

$$S_C(e_i, e_j) = \frac{e_i^\top e_j}{\|e_i\|_2 \cdot \|e_j\|_2} \tag{9}$$

식 (9)를 통해 산출된 코사인 유사도를 활용하여, 제안하는 파이프라인은 그림 6과 같이 크게 두 단계의 객체 재식별 후처리를 수행한다. 첫째는 가림 현상 등으로 인해 단절되었던 추적 ID를 하나로 통합하는 전역 ID(Global ID) 할당 과정이며, 둘째는 프레임 간 객체 구성의 일치도를 평가하여 의미론적 중복 프레임을 제거하는 과정이다. 각 과정의 구체적인 동작 방식은 이어지는 소절에 기술되어 있다.

6.1 전역 ID 할당(Global ID Assignment)

전역 ID 할당 단계에서는 동일 클래스 내 객체들을 코사인 유사도 임계값을 기준으로 군집화(Clustering)하여 전역 ID를 부여한다. 구체적으로, 각 클래스별로 모든 객체 쌍에 대해 산출된 코사인 유사도가 사전에 정의된 재식별 허용 임계값 이상인 경우 이를 동일 객체로 판정한다. 이후 코사인 유사도 임계값 기반의 탐욕적 군집화(Greedy clustering)를 수행하여, 클래스명과 군집 번호가 결합된 형태의 새로운 전역 ID를 할당한다. 이 과정을 통해 가림 현상이나 프레임 누락 등으로 인해 서로 다른 추적 ID(Track ID)가 부여되었던 동일 객체가 파이프라인 전체에서 일관된 단일 식별자로 성공적으로 통합된다.

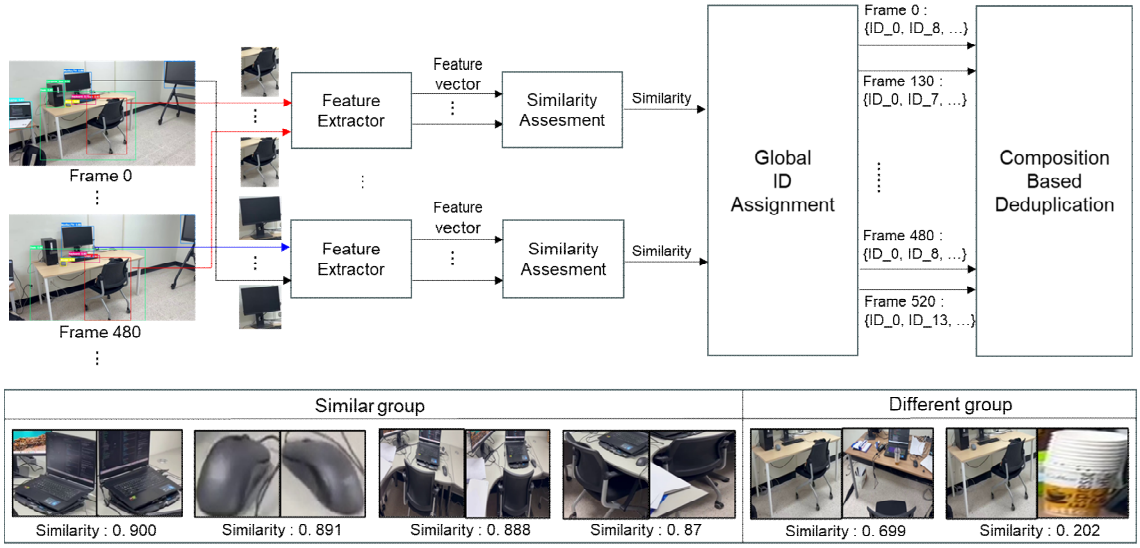

그림 6. OSNet Re-ID 기반 Global ID 할당 및 객체 구성 유사도 기반 프레임 중복 제거
Fig. 6. Global ID assignment via OSNet Re-ID and object composition-based frame deduplication

6.2 객체 구성 기반 프레임 중복 제거

전역 ID 부여가 완료된 이후에는 프레임 내 객체 구성을 비교하여 의미론적 중복 프레임을 최종적으로 제거한다. 임의의 두 프레임 쌍이 중복 프레임으로 판정되어 정보량이 적은 프레임이 제거되기 위해서는 다음의 세 가지 조건을 모두 만족해야 한다. 첫째, 두 프레임 내에서 탐지된 객체의 총 개수가 동일해야 한다. 둘째, 전역 ID를 기준으로 두 프레임 간에 공통으로 존재하는 객체가 적어도 하나 이상 있어야 한다. 셋째, 이러한 공통 객체들 중에서 재식별 코사인 유사도가 사전에 설정된 병합 임계값(0.7)을 초과하는 쌍의 비율이 전체 탐지 객체 수의 95%를 초과해야 한다.

첫 번째 조건에 의해 두 프레임의 탐지 객체 수가 완벽히 동일하도록 강제되므로, 이는 거의 모든 등장 객체가 매우 높은 수준의 외관 유사도를 공유하는 의미론적(Semantic) 중복 장면만을 선별적으로 제거하는 매우 엄격한 기준이 된다. 이 단계는 앞선 해싱 및 히스토그램 필터링 단계에서 걸러지지 않은 고차원적이고 의미론적인 중복 프레임들을 확실하게 배제함으로써, 최종 단계인 탐욕적 키프레임 선택 알고리즘으로 전달되는 후보군의 품질을 극대화하는 역할을 수행한다.

7. 탐욕적 키프레임 선택(Greedy Coverage Selection)

최종 키프레임 선택은 영상 내에 등장하는 다양한 객체들의 상호작용 및 구성 패턴을 최대한 많이 포함하기 위한 최대 커버리지 문제(Maximum coverage problem)로 정의된다. 이를 위해 먼저 객체 재식별(Re-ID) 단계를 거쳐 통합된 전역 ID를 바탕으로, 각 프레임 내 존재하는 객체들의 k-조합(k-combination) 집합을 추출한다. 구체적인 조합 집합 추출 방식은 식 (10)과 같다.

$$C_t = \{(o_i, o_j) \mid o_i, o_j \in O_t,\ i < j\}, \quad k = 2 \tag{10}$$

식 (10)에서 특정 프레임의 조합 집합은 해당 프레임에 포함된 전체 객체 집합 내에서 서로 다른 두 객체(k=2)로 구성될 수 있는 모든 쌍(Pair)들의 모음을 의미한다. 본 연구에서는 단일 객체나 세 개 이상의 조합보다, 두 객체 간의 관계가 영상의 주요 장면을 구성하는 핵심 단위라고 판단하여 k를 2로 설정하였다. 이렇게 추출된 객체 조합들을 바탕으로, 기존에 선택된 키프레임들이 커버하지 못한 새로운 조합을 가장 많이 포함하고 있는 프레임을 반복적으로

우선 선별하는 탐욕적(Greedy) 알고리즘을 적용한다. 이와 관련된 전체적인 키프레임 선택 동작 절차는 알고리즘 2에 상세히 기술되어 있다.

알고리즘 2. k-조합 기반 탐욕적 키프레임 선택
Algorithm 2. Greedy Keyframe Selection via k-Combinations

Algorithm 2 Greedy Keyframe Selection via k-Combinations

Require: Filtered frames $F = \{f_1, \ldots, f_m\}$, combination size k, max selection $N_{\max}$
Ensure: Selected keyframes K
$S \leftarrow \emptyset; \quad K \leftarrow \emptyset$
for each $f_t \in F$ **do**
  $C_t \leftarrow \{(o_i, o_j) \mid o_i, o_j \in O_t,\ i < j\}$ ▷ k-combinations, Eq. (10)
end for
while $|K| < N_{\max}$ **and** $\exists f_t : C_t \setminus S \neq \emptyset$ **do**
  $f^* \leftarrow \arg\max_t |C_t \setminus S|$
  $K \leftarrow K \cup \{f^*\}; \quad S \leftarrow S \cup C_{f^*}$
end while
return K

알고리즘 2에서 설명된 바와 같이 매 반복마다 최대의 정보 이득을 주는 프레임을 선택하는 탐욕적 접근법은 하위 모듈러(Submodular) 함수 최대화 문제의 전형적인 특성을 가진다. 따라서 이 방식은 다항 시간 내에 효율적으로 해를 도출하면서도, 글로벌 최적해 대비 최소 (1 - 1/e), 즉 약 63.2% 이상의 근사비(Approximation ratio)를 수학적으로 보장하는 장점이 있다[25].

8. 프레임 선별 후처리 및 깊이 정보 추출

탐욕적 키프레임 선택 이후, 최종 선별된 키프레임 집합 내에 잔존할 수 있는 미세한 시각적 중복을 배제하기 위해 2차 히스토그램 필터링(Post-Greedy Profile Tracking)을 수행한다. 이 과정은 앞선 1차 필터링 단계와 동일하게 바타차야(Bhattacharyya) 거리 기반의 양방향 슬라이딩 윈도우 방식을 사용하지만, 1차 필터링 시점보다 훨씬 더 엄격한 수준의 거리 허용 임계값을 적용함으로써 최종 결과물의 시각적 다양성을 극대화한다.

중복 제거가 완전히 마무리된 최종 키프레임 집합에 대해서는, 그림 7에 나타난 바와 같이 DPT(Dense Prediction Transformer) 단안 깊이 추정 모델을 적용하여 각 프레임의 공간적 깊이(Depth) 맵을 추가로 추출한다[15]. 추출된 깊이 맵은 2D 이미지 내 각 픽셀의 상대적 카메라 거리를 정량적인 수치로 나타낸다. 제안하는 시스템은 단순히 프레임 전체의 깊이를 구하는 것에 그치지 않고, 탐지된 각 객체의 바운딩 박스 내부 영역에 대해 Otsu 임계화 기법을 적용

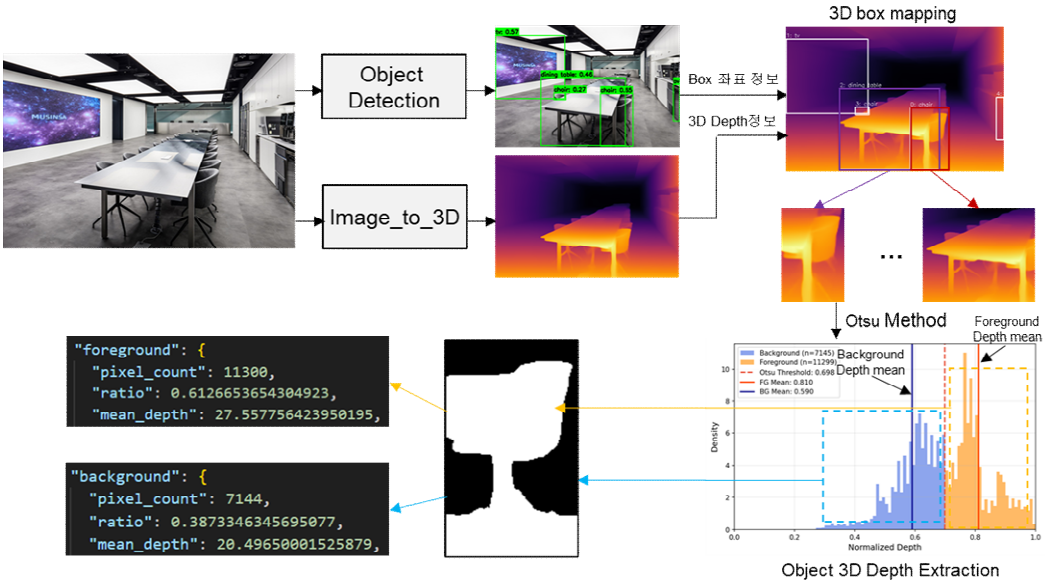

그림 7. DPT 기반 단안 깊이 추정 및 Otsu 임계화를 통한 객체 전경 분리
Fig. 7. Monocular depth estimation via DPT and foreground segmentation using Otsu thresholding

하여 객체에 해당하는 전경(Foreground)과 배경(Background) 픽셀을 자동적이고 정밀하게 분리한다.

그림 7의 하단 분포도에서 볼 수 있듯이, 임계값을 기준으로 분리된 전경 픽셀들만을 대상으로 깊이 값의 평균, 표준편차 및 주요 백분위수(25/50/75 백분위)를 산출한다. 이를 통해 기존 2D 기반 키프레임이 가지는 공간 정보의 부재를 극복하고, 각 등장 객체의 3차원적 위치 정보를 포함하는 고품질의 최종 메타데이터를 완성한다.

Ⅳ. 실험 및 결과

1. 실험 설정

본 연구의 실험은 NVIDIA RTX 4070 Ti GPU와 Intel Core i7 프로세서가 탑재된 하드웨어 환경에서 수행되었다. 소프트웨어 구현에는 Python 3.10, PyTorch 2.0 및 Ultralytics 프레임워크를 사용하였으며, 객체 탐지 모델로는 YOLOv26을 채택하였다. 제안하는 파이프라인의 성능 평가를 위한 데이터셋으로는 비디오 요약 분야의 대표적인 공개 벤치마크인 TVSum을 활용하였다[21]. 표 1은 본 실험에 사용된 TVSum 데이터셋의 세부 통계를 나타낸다.

TVSum 데이터셋은 뉴스, 하우투(how-to), 다큐멘터리, 브이로그, 1인칭 시점 등 10개 카테고리의 유튜브 비디오 50편으로 구성되어 있다. 해당 비디오들은 평균 약 250초의 길이를 가지며, 일상적 객체의 빈번한 등·퇴장, 동적 카메라 움직임, 다양한 조명 변화 및 가림 현상(occlusion)을 포함한다. 이러한 영상의 복잡성은 다중 객체 추적(MOT) 과정에서 필연적으로 발생하는 ID 단절 문제와 재식별(Re-ID) 파이프라인의 연산 병목 현상을 검증하기 위한 극한 테스트 환경(stress test)을 제공하므로, 제안 알고리즘의 강건성 및 범용성 검증에 적합하다. 다만, TVSum 데이터셋이 기본 제공하는 크라우드소싱 기반 정답(ground truth, GT) 데이터를 본 연구의 평가 지표로 직접 활용하는 데에는 두 가지 근본적 한계가 존재한다. 첫째, 태스크 정의의 불일치(task mismatch) 문제이다. 기존 TVSum의 GT는 시청자가 주관적으로 느끼는 '흥미도'를 기준으로 산정되어 일반적 비디오 요약을 목적으로 설계되었다. 반면, 본 연구의 핵심 목표는 시청 흥미도가 아닌 영상 내 객체 구성의 물리적 변화(신규 객체 등장, 기존 객체 퇴장, 명확한 장면 전환 등)를 대표하는 키프레임의 객관적 선별에 있으므로, 측정 대상 및 평가 기준이 본질적으로 상이하다. 둘째, 주관성에 기인한 평가자 간 편차 문제이다. 선행 연구[22]에 따르면, TVSum은 비디오 수가 제한적이어서 특정 토픽에 대한 편향(topic bias)이 발생하기 쉬우며, 요약의 본질적 주관성으로 인해 평가자 간 점수를 단순 평균하여 단일 GT로 활용하는 방식은 신뢰성 있는 일반화 평가를 저해한다는 지적이 제기된 바 있다[23]. 이러한 한계를 극복하기 위해 본 연구에서는 TVSum의 원본 비디오 50편 전체를 활용하여 데이터의 다양성을 유지하되, 본 연구의 태스크 정의에 부합하는 새로운 GT를 독립적으로 구축하였다. GT 구축 과정은 복수의 평가자가 각 비디오를 독립적으로 시청하며 (1) 신규 객체의 등장, (2) 기존 객체의 퇴장으로 인한 객체 구성 변화, (3) 장면 전환에 해당하는 프레임을 식별한 후,

표 1. 실험 데이터셋 통계
Table 1. Statistics of the experimental dataset

VideoID	Category	Duration(s)	#Frames	#GT_KFs
AwmHb44_ouw	Changing Vehicle Tire	354	10597	37
HT5vyqe0Xaw	Getting Vehicle Unstuck	322	9671	25
i3wAGJaaktw	Grooming an Animal	156	4700	7
WG0MBPpPC6I	Making Sandwich	397	9535	29
91IHQYk1IQM	Parade	110	3312	28
...	...	...	...	...
Avg.(50 videos)	-	251.3	7047.1	29.3

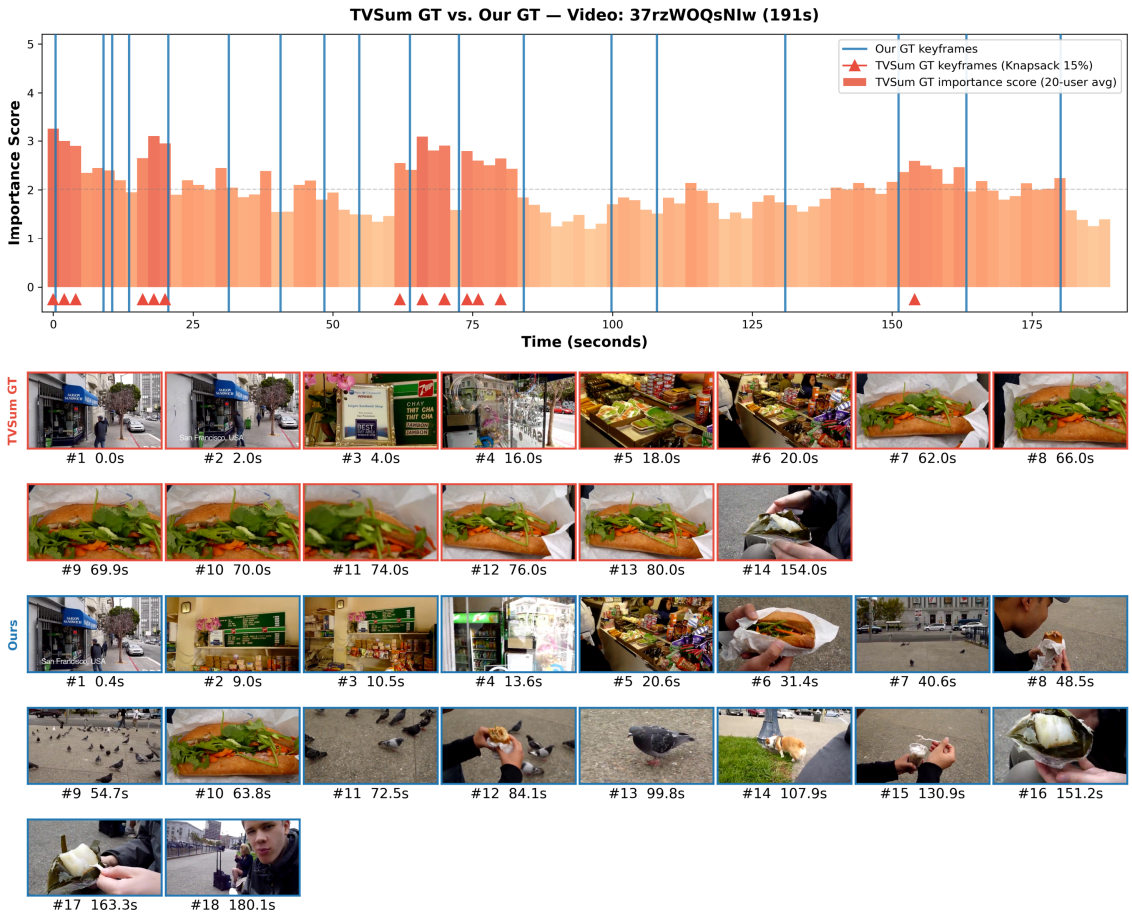

그림 8. 기존 TVSum GT와 제안 GT의 비교(비디오: 37rzWOQsNIw, 191초)
Fig. 8. Comparison between original TVSum GT and proposed GT(Video: 37rzWOQsNIw, 191s)

다수결 합의(majority voting)를 통해 최종 키프레임을 확정하였다.

그림 8은 TVSum 비디오 '37rzWOQsNIw'(Making Sandwich, 191초)를 대상으로 기존 TVSum GT와 본 연구에서 구축한 GT의 차이를 시각적으로 비교하였다. 그림 8 상단의 시간축 그래프에서, 빨간색 막대 그래프는 TVSum 원본 GT의 20명 평균 흥미도 점수 분포를 나타내며, 빨간 삼각형(▲)은 해당 점수에 기반하여 Knapsack 알고리즘(예산 15%)으로 선택된 TVSum GT 키프레임 14개의 위치를, 파란 수직선은 본 연구의 제안 GT 키프레임 18개의 위치를 각각 나타낸다. TVSum GT 키프레임은 0~6초 구간에 3개, 16~22초 구간에 3개, 62~80초 구간에 7개가 집중되어, 전체 14개 중 13개(93%)가 영상 전반부 80초 이내에 밀집되어 있다. 반면, 80~191초에 해당하는 후반부 58%의 구간에는 단 1개의 키프레임만이 배치되었다.

이에 반해 제안 GT 키프레임은 0.4초부터 180.1초까지 전체 영상에 걸쳐 평균 10.6초 간격으로 균일하게 분포하고 있다. 그림 8 하단의 프레임 비교에서 TVSum GT가 선택한 프레임들은 샌드위치의 클로즈업 장면(#7~#13, 62~80초)이 반복적으로 나타나며, 영상 후반부에 등장하는 야외 장면 및 객체 변화를 포착하지 못한다. 반면, 제안 GT의 프레임들은 가게 외관(#1), 매장 내부 진열대(#2~#3), 샌드위치 재료 및 제작 과정(#4~#5), 포장(#6), 야외 이동(#7~#9), 비둘기 등 새로운 객체의 등장(#11~#14), 최종 식

$$\mathrm{Precision}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}, \quad \mathrm{Recall}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}, \quad \mathrm{F1}_i = 2 \times \frac{\mathrm{Precision}_i \times \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i} \tag{11}$$

$$\mathrm{P}_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Precision}_i, \quad \mathrm{R}_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Recall}_i, \quad \mathrm{F1}_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{F1}_i \tag{12}$$

사 장면(#17~#18)에 이르기까지 영상 내 객체 구성의 변화를 빠짐없이 포착하고 있다. 이는 제안 GT가 특정 하이라이트 구간에 편중되지 않고 영상 전체의 의미론적 변화를 균형 있게 반영함을 정성적으로 입증한다.

제안 파이프라인의 성능 검증을 위한 정량적 평가 지표로는 정밀도(Precision), 재현율(Recall), F1-score를 채택하였다. 모델이 추출한 키프레임과 GT 간의 시간적 매칭 허용 오차는 ±15프레임(30fps 기준 약 0.5초)으로 설정하였다. 또한, 처리 효율성의 다각적 검증을 위해 선별 프레임 수, 객체 재식별(Re-ID) 소요 시간, 전체 파이프라인 처리 시간을 함께 측정하였다. 각 지표의 산출 수식은 식 (11)과 같다.

식 (11)에서 i는 개별 비디오를 나타내며, TP_i, FP_i, FN_i는 각각 i번째 비디오에서의 진양성(True Positive), 위양성(False Positive), 위음성(False Negative) 키프레임 수이다. 비디오 길이 편차에 의한 편향을 방지하기 위해, 본 연구에서는 각 비디오별로 Precision, Recall, F1-score를 개별 산출한 후 전체 N편에 대해 평균을 취하는 Macro-average 방식을 채택하였으며, 이는 식 (12)와 같다.

따라서 표 2에 보고된 F1-score(0.7077)는 전체 Precision(0.8103)과 Recall(0.6520)의 직접 조화평균(0.7225)과 다를 수 있으며, 이는 개별 비디오 단위로 F1을 먼저 계산한 후 평균하는 Macro-average 특성에 의한 정상적인 차이이다.

성능 비교를 위한 대조군으로는 일정 시간 간격으로 프레임을 기계적으로 추출하는 균일 샘플링(Uniform Sampling, interval=5s 및 15s) 방식, 기존 추적 모델의 단순 결합 구성인 YOLO+BoTSORT+Re-ID 기반 파이프라인, 그리고 이에 탐욕적 선택 알고리즘을 추가한 구성(YOLO+BoTSORT+Re-ID+Greedy)을 설정하여 제안 방법과의 비교 평가를 수행하였다.

2. 정량적 결과

제안 방법과 비교 방법들의 정량적 성능(Precision, Recall, F1-score) 및 처리 효율(선별 프레임 수, 연산 시간)은 표 2와 표 3에서 볼 수 있고, 모든 결과는 TVSum 데이터셋 전체 50편의 평균값으로 계산하여 비교하였다.

제안 방법의 유효성을 검증하기 위해, 먼저 기계적 추출 방식인 균일 샘플링(Uniform Sampling)과의 비교 실험을 수행하였다. 균일 샘플링은 영상의 전체 길이와 무관하게 5초(interval=5s) 및 15초(interval=15s)의 고정 시간 간격으로 프레임을 추출하는 방식으로, 두 설정 모두 F1-score 0.15 수준의 저조한 성능을 나타내었다. 이러한 결과는 고정 간격 샘플링이 지정 구간 사이에 발생하는 객체의 등·퇴장 및 장면 전환 등 의미론적(semantic) 변화를 포착하지

표 2. 정량적 성능 비교(TVSum 50편 평균)
Table 2. Quantitative performance comparison(averaged over 50 TVSum videos)

Method	Precision	Recall	F-1 score
Uniform Sampling (interval=5s)	0.1314	0.1982	0.1502
Uniform Sampling (interval=15s)	0.1343	0.0675	0.0858
YOLO + BoT-SORT + Re-ID	0.0652	0.8380	0.1142
YOLO + BoT-SORT + Re-ID + Greedy	0.1653	0.7152	0.2450
Proposed (Ours)	**0.8103**	**0.6520**	**0.7077**

표 3. 처리 효율 비교(TVSum 50편 평균)
Table 3. Processing efficiency comparison(averaged over 50 TVSum videos)

Method	#Frames	ReID (s)	Total (sec)
Uniform Sampling (interval=5s)	50.78	-	0.0231
Uniform Sampling (interval=15s)	17.28	-	0.0045
YOLO + BoT-SORT + Re-ID	803.94	74.8822	187.9174
YOLO + BoT-SORT + Re-ID + Greedy	227.52	74.7738	172.973
Proposed (Ours)	**26.84**	**1.481**	**121.4426**

못해 재현율(Recall)이 낮아지는 한편, 정적 구간이 장시간 지속될 경우 동일 장면의 중복 추출로 인해 정밀도(Precision) 역시 저하되는 구조적 한계에 기인한다.

YOLO 검출기와 BoTSORT 추적기, OSNet 기반 재식별(Re-ID)을 단순 결합한 추적 기반 파이프라인(YOLO+BoTSORT+Re-ID)은 0.8380의 높은 재현율을 달성하였으나, 정밀도는 0.0652에 그쳤다. 이는 추적기 고유의 한계로 인한 추적 ID(Track ID) 단절 현상에 주로 기인한다. 가림(occlusion) 및 객체 간 교차 상황에서 동일 객체에 신규 ID가 반복적으로 할당됨에 따라, 실제로는 동일한 장면임에도 객체 구성이 변경된 것으로 오인되어 불필요한 프레임 그룹이 다수 생성되었다. 그 결과, 영상당 평균 803.94장의 프레임이 추출되어 오탐(false positive)이 현저히 증가하였다. 탐욕적 선택 알고리즘을 추가한 구성(YOLO+BoTSORT+Re-ID+Greedy)에서는 선별 프레임 수가 227.52장으로 감소하고 정밀도가 0.1653으로 소폭 개선되었으나, 입력 프레임 후보군 자체에 이미 ID 단절 오류가 내재되어 있어 근본적인 성능 개선에는 한계가 존재하였다.

이에 반해 제안 방법은 정밀도 0.8103, 재현율 0.6520, F1-score 0.7077을 달성하여, 비교군 최고 성능(0.2450) 대비 약 2.88배의 향상을 보였다. 이러한 성능 개선은 제안 파이프라인 전반에 걸쳐 유기적으로 통합된 다단계 중복 제거 및 보정 메커니즘에 기인한다. 구체적으로, 전처리 단계에서 Crop Histogram 기반 슬라이딩 윈도우 매칭을 통해 가림 현상으로 단절된 객체 ID를 선제적으로 복구하고, 하이브리드 연속성 점수(Hybrid Continuity Score)를 도입하여 불안정한 ID 변화에도 강건한 프레임 그룹 유지가 가능하도록 설계하였다. 후처리 단계에서는 dHash 및 바타차야(Bhattacharyya) 거리 기반의 이중 시각 필터링을 적용하여 구조적·색상 기반 중복을 효과적으로 제거함으로써, 오탐을 대폭 억제하고 영상의 핵심 내용을 대표하는 진양성(true positive) 프레임을 선별할 수 있었다.

표 3에 제시된 처리 효율 비교 결과는 제안 방법이 정확도 향상뿐만 아니라 연산 효율 측면에서도 기존 방식의 병목(bottleneck) 현상을 효과적으로 완화하였음을 보여준다. 객체 재식별(Re-ID) 단계의 처리 시간을 구체적으로 분석하면, 기존 방식에서는 약 800여 장에 달하는 후보 프레임 전체를 고연산량의 OSNet 모델에 입력한 후 코사인 유사도 기반의 행렬 연산을 수행하므로, 평균 74.88초의 상당한 연산 지연이 발생하였다. 반면, 제안 파이프라인에서는 연산 부담이 상대적으로 낮은 dHash 및 히스토그램 기반 필터링 모듈을 Re-ID 단계 이전에 전진 배치하여, Re-ID 모델에 입력되는 후보 프레임 수를 영상당 평균 26.84장 수준으로 대폭 축소하였다. 그 결과, Re-ID 처리 시간이 1.48초로 약 50배 단축되었으며, 전체 시스템의 총 처리 시간 역시 187.92초에서 121.44초로 약 35% 감소하여 성능 향상과 연산 효율성을 동시에 확보하였다.

3. 추가 비교 실험(Ablation Study)

제안 파이프라인을 구성하는 각 모듈의 개별 기여도를 정량적으로 검증하기 위해 표 4와 같이 구성 요소를 순차적으로 추가하는 절제 실험을 수행하였다. 실험 구성 (a) Baseline은 프레임 건너뛰기 간격(frame_skip)을 5로 설정하고, 추적 ID 집합의 완전 일치 여부만을 기준으로 그룹을 분리하며, 각 그룹 내 시간적 중앙 프레임을 기계적으로 선택하는 기본 구성에 해당한다. 단, 탐욕적(greedy) k-조합 기반 키프레임 선택 알고리즘은 모든 실험 구성에 공통으

표 4. Ablation Study 결과(TVSum 50편 평균)
Table 4. Ablation study results(averaged over 50 TVSum videos)

Test Case	Skip	Crop Hist.	Hybrid Cs	Best-Q	dHash	profile	Post-filter	Prec.	Rec.	F1	Avg. frames	Total (sec)
Baseline	O	X	X	X	X	X	X	0.237	0.488	0.283	96.54	39.55
+All-Frames	X	X	X	X	X	X	X	0.167	0.713	0.246	227.3	168.9
+Crop Hist.	X	O	X	X	X	X	X	0.167	0.713	0.246	227.3	197.3
+Hybrid CS	X	O	O	X	X	X	X	0.325	0.200	0.240	20.34	117.2
+Best Quality	X	O	O	O	X	X	X	0.783	0.648	0.694	27.64	118.0
+dHash	X	O	O	O	O	X	X	0.798	0.650	0.701	27.18	117.8
+Profile	X	O	O	O	O	O	X	0.810	0.652	0.708	26.84	117.7
+Full (ours)	O	O	O	O	O	O	O	**0.810**	**0.652**	**0.708**	**26.84**	**117.8**

로 포함되어 있다.

표 4의 수치 변화를 분석한 결과, 파이프라인 각 단계의 역할과 인과관계가 명확히 확인된다. 우선, (a) Baseline에서 프레임 스킵을 제거한 (b) +All-Frames 구성을 적용한 경우, 탐지 누락이 방지되어 재현율이 0.488에서 0.713으로 현저히 상승하였다. 그러나 기존 추적기 고유의 ID 단절 문제로 인해 동일 장면이 다수의 그룹으로 과도하게 분할되면서, 선별 프레임 수가 96.54장에서 227.3장으로 급증하고 정밀도는 0.237에서 0.167로 하락하는 상충 관계(trade-off)가 관찰되었다. 이러한 과분할 문제는 (c) Crop Hist. 모듈 및 (d) Hybrid CS의 순차적 추가를 통해 해소되었다. 시각적 히스토그램 기반의 단절 ID 복구와 다중 지표를 결합한 소프트 그룹핑이 적용됨에 따라, 후보 프레임 수가 227.3장에서 20.34장으로 대폭 감소하여 불필요한 프레임 분할이 효과적으로 억제되었다. 성능 지표상 가장 현저한 개선은 (e) Best-Quality 모듈 추가 단계에서 나타났다. 이는 그룹의 시간적 중앙 프레임을 일률적으로 선택하던 기존 방식에서 탈피하여, 그룹 내 탐지 객체 수 및 신뢰도가 가장 높은 고품질 프레임을 대표 프레임으로 선정하도록 변경한 결과이다. 이를 통해 모션 블러 또는 가림 현상이 심한 저품질 프레임이 배제되면서 정밀도가 0.325에서 0.783으로 급격히 향상되었다. 이후 (f) dHash 및 (g) Profile Tracking을

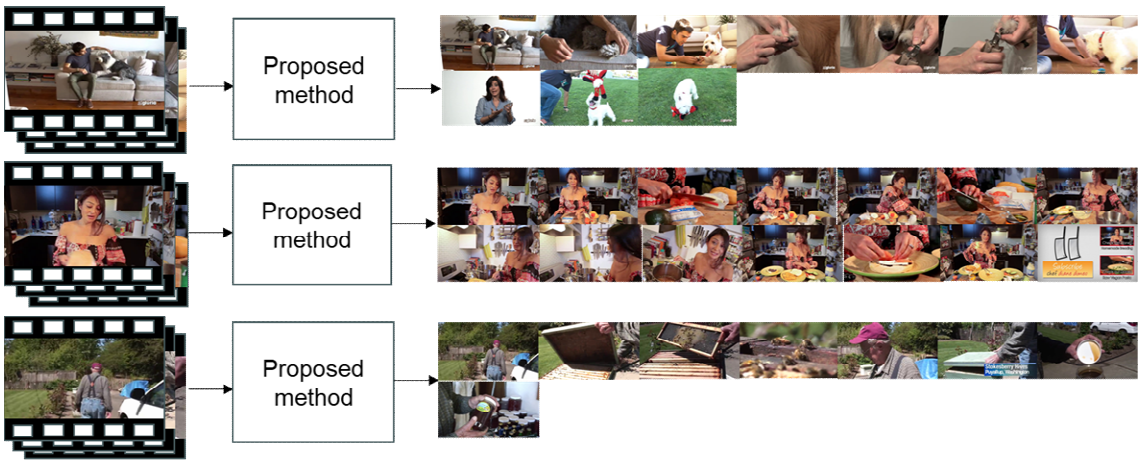

그림 9. 제안 파이프라인의 최종 키프레임 추출 결과 예시
Fig. 9. Example of final keyframes extracted by the proposed pipeline

순차적으로 적용하여 미세한 조명 변화와 구조적·시각적 중복을 추가로 제거함으로써 정밀도를 0.810까지 향상시켰다.

최종적으로 (h) Post-filter 단계를 거쳐 선별 프레임 수를 26.84장으로 최적화하는 동시에, 전체 실험 구성 중 최고 수준의 F1-score(0.708)를 달성하였다. 이러한 결과는 본 연구에서 제안하는 다단계 모듈들이 단순한 프레임 수 감소에 그치지 않고 상호 보완적으로 작용하여, 키프레임의 대표성과 처리 효율 간 최적의 균형점을 확보하고 있음을 실증적으로 입증한다.

상기 정량적 성능 향상은 실제 추출된 키프레임의 시각적 분석을 통해서도 확인된다. 그림 9에는 제안 방법이 다양한 비디오 시퀀스로부터 핵심 장면을 선별하는 과정의 추출 결과 예시가 제시되어 있다. 선별된 키프레임 집합을 관찰한 결과, 영상 내 등장하는 다양한 객체 조합 및 장면 구성을 포괄하면서도 시각적으로 유사한 중복 프레임은 효과적으로 배제되어 있음을 확인할 수 있다. 특히 신규 객체의 최초 등장 시점 또는 기존 객체의 퇴장으로 인해 전체 장면의 맥락이 전환되는 시점이 핵심 프레임으로 정확히 포착되었다. 이는 제안 파이프라인이 단순한 픽셀 수준의 화면 변화가 아닌, 고수준의 의미론적 객체 구성 변화를 충실히 반영하여 영상 요약을 생성함을 정성적으로 검증한다.

V. 결 론

본 연구에서는 다중 객체 추적(MOT)과 객체 재식별(Re-ID)을 유기적으로 통합하여 비디오 내 핵심 의미를 포착하는 키프레임 추출 파이프라인을 제안하였다. 제안 파이프라인은 다중 모델의 단순 결합 방식에서 빈번히 발생하는 추적 ID(Track ID) 단절로 인한 오탐 증가 및 이에 수반되는 연산 병목 현상을 효과적으로 개선하였다. TVSum 데이터셋 50편을 대상으로 한 종합 평가 결과, 제안 방법은 정밀도 0.8103, 재현율 0.6520, F1-score 0.7077을 달성하여, 본 연구에서 비교를 위해 구축한 대조군(YOLO+BoTSORT+Re-ID+Greedy, F1-score 0.2450) 대비 약 2.89배의 성능 향상을 기록하였다. 연산 효율성 측면에서도 고연산량 재식별 모델의 입력 프레임 수를 대조군의 평균 803.94장에서 26.84장으로 대폭 축소함으로써, 전체 처리 시간을 187.92초에서 121.44초로 약 35% 단축하였다. 추가 비교 실험(Ablation Study)을 통해 제안 파이프라인 내 각 모듈의 개별 기여도를 검증하였다. 기계적 프레임 스킵을 배제한 전 프레임 탐지를 통해 탐지 누락을 방지하는 한편, Crop Histogram 기반 Track ID 복원 및 Hybrid Continuity Score 도입을 통해 프레임 과분할 문제를 효과적으로 억제하였다. 특히, 그룹 내 탐지 객체 수와 신뢰도를 기준으로 최적 프레임(Best-Quality)을 선택하는 전략이 오탐 감소 및 정밀도 향상에 핵심적 역할을 수행함을 확인하였다. 이후 dHash 기반 구조적 중복 제거와 Profile Tracking 기반 시각적 필터링을 순차 적용하여, 최종 선별 프레임 수를 26.84장으로 압축하면서도 전체 실험 구성 중 최고 F1-score(0.708)를 달성함으로써 키프레임의 대표성과 처리 효율 간 균형을 확보하였다. 아울러, 선별된 2D 키프레임에 DPT 기반 단안 깊이 추정을 적용하여 각 객체의 3차원 공간 좌표 및 전경 분리 통계를 포함한 확장 메타데이터를 생성함으로써, 3D 재구성 등 다양한 하류 태스크(downstream task)로의 확장 가능성을 제시하였다. 본 연구의 한계점으로는 다음 사항이 존재한다. 첫째, 극단적 조명 변화 또는 심각한 가림 현상(occlusion) 발생 시 Crop Histogram 매칭의 신뢰도가 저하되어 유사 외형의 상이한 객체를 오인할 가능성이 있다. 둘째, dHash의 중복 판정 임계값이 고정되어 있어 영상 특성에 따른 유연한 대응이 제한된다. 따라서 향후 연구에서는 이러한 파라미터를 영상의 동적 특성에 따라 적응적(adaptive)으로 조절하는 기법을 탐구하고자 한다. 또한, 파이프라인의 실시간 처리를 위한 모듈 간 병렬 연산 최적화와 함께, 보행자 외 다양한 도메인으로의 확장을 위한 재식별 모델의 도메인 적응(domain adaptation) 기법에 관한 후속 연구를 진행할 계획이다.

참 고 문 헌 (References)

[1] S. Kaur et al., “An effective Key Frame Extraction technique based on Feature Fusion and Fuzzy-C means clustering with Artificial Hummingbird,” Scientific Reports, Vol. 14, No. 1, p. 26651, 2024. doi: https://doi.org/10.1038/s41598-024-75923-y

[2] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection," arXiv preprint arXiv:2509.25164, 2025. doi: https://doi.org/10.48550/arXiv.2509.25164

[3] K. Zhou et al., "Omni-Scale Feature Learning for Person Re-Identification," Proc. IEEE/CVF ICCV, pp. 3702-3712, 2019. doi: https://doi.org/10.1109/ICCV.2019.00380

[4] Y. Zhang et al., "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," Proc. ECCV, pp. 201-217, 2022. doi: https://doi.org/10.1007/978-3-031-20047-2_1

[5] N. Wojke et al., "Simple Online and Realtime Tracking with a Deep Association Metric," Proc. IEEE ICIP, pp. 3645-3649, 2017. doi: https://doi.org/10.1109/ICIP.2017.8296962

[6] A. Bewley et al., "Simple Online and Realtime Tracking," Proc. IEEE ICIP, pp. 3464-3468, 2016. doi: https://doi.org/10.1109/ICIP.2016.7533003

[7] S. Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE TPAMI, Vol. 39, No. 6, pp. 1137-1149, 2017. doi: https://doi.org/10.1109/TPAMI.2016.2577031

[8] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," Proc. IEEE CVPR, pp. 779-788, 2016. doi: https://doi.org/10.1109/CVPR.2016.91

[9] M. Luo et al., "A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification," IEEE TMM, Vol. 22, No. 10, pp. 2597-2609, 2020. doi: https://doi.org.10.1109/TMM.2019.2958756

[10] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," Proc. IEEE, Vol. 109, No. 11, pp. 1838-1863, 2021. doi: https://doi.org/10.1109/JPROC.2021.3117472

[11] J. Schönberger and J. Frahm, "Structure-from-Motion Revisited," Proc. IEEE CVPR, pp. 4104-4113, 2016. doi: https://doi.org/10.1109/CVPR.2016.445

[12] Z. Ge, et al., "YOLOX: Exceeding YOLO Series in 2021," arXiv preprint arXiv:2107.08430, 2021. doi: https://doi.org/10.48550/arXiv.2107.08430

[13] L. Zheng et al., "Scalable Person Re-identification: A Benchmark," Proc. IEEE ICCV, pp. 1116-1124, 2015. doi: https://doi.org/10.1109/ICCV.2015.133

[14] B. Truong and S. Venkatesh, "Video Abstraction: A Systematic Review," ACM TOMM, Vol. 3, No. 1, pp. 1-37, 2007. doi: https://doi.org/10.1145/1198302.1198305

[15] R. Ranftl et al., "Vision Transformers for Dense Prediction," Proc. IEEE/CVF ICCV, pp. 12179-12188, 2021. doi: https://doi.org/10.1109/ICCV48922.2021.01196

[16] R. Ranftl et al., "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer," IEEE TPAMI, Vol. 44, No. 3, pp. 1623-1637, 2022. doi: https://doi.org/10.1109/TPAMI.2020.3019967

[17] L. Yang et al., "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," Proc. IEEE/CVF CVPR, 2024. doi: https://doi.org/10.1109/CVPR52733.2024.00987

[18] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed., Cambridge Univ. Press, 2004. doi: https://doi.org/10.1017/CBO9780511811685

[19] C. Wu, "Towards Linear-Time Incremental Structure from Motion," Proc. 3DV, pp. 127-134, 2013. doi: https://doi.org/10.1109/3DV.2013.25

[20] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," arXiv preprint arXiv:2206.14651, 2022. doi: https://doi.org/10.48550/arXiv.2206.14651

[21] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing Web Videos Using Titles," Proc. IEEE/CVF CVPR, pp. 5179-5187, 2015. doi: https://doi.org/10.1109/CVPR.2015.7299154

[22] J. Sul, J. Han, and J. Lee, "Mr. HiSum: A Large-scale Dataset for Video Highlight Detection and Summarization," Advances in Neural Information Processing Systems, Vol. 36, 2023. doi: https://doi.org/10.52202/075280-1764

[23] Z. Zhou et al., "Video Summarization using Denoising Diffusion Probabilistic Model," Proc. AAAI, 2025. doi: https://doi.org/10.1609/aaai.v39i7.32727

[24] Y. Zhao et al., "DETRs Beat YOLOs on Real-time Object Detection," Proc. IEEE/CVF CVPR, pp. 16965-16974, 2024. doi: https://doi.org/10.1109/CVPR52733.2024.01605

[25] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An Analysis of Approximations for Maximizing Submodular Set Functions," Mathematical Programming, Vol. 14, No. 1, pp. 265-294, 1978. doi: https://doi.org/10.1007/BF01588971

저 자 소 개

이 승 복

- 2021년 ~ 현재 : 세종대학교 인공지능학부 학사과정
- ORCID : https://orcid.org/0009-0002-3531-3472
- 주관심분야 : 컴퓨터 비젼, 멀티모달 AI, 비디오 요약 및 캡셔닝

민 병 석

- 2001년 : 연세대학교 전기공학 (학사)
- 2003년 : 연세대학교 전기전자공학과 (석사)
- 2009년 : 미국 Purdue대학교 전기컴퓨터공학과 (박사)
- 2009년 ~ 2020년 : 삼성전자 영상디스플레이 사업부 수석연구원
- 2020년 ~ 2024년 : ㈜ 자비스 연구소장
- 2024년 ~ 2025년 : ㈜ 마크애니 연구소장
- 2025년 ~ 현재 : 세종대학교 인공지능데이터사이언스학과 교수
- ORCID : https://orcid.org/0000-0001-9826-3471
- 주관심분야 : 컴퓨터 비젼 및 영상처리