Diffusion 모델 기반 깊이 추정을 통한 K-POP 직캠 인물 검출 성능 향상

송준호, 라윤경, 박범준, 최한비, *홍민수
서울시립대학교, *한국방송공사
1999junho@uos.ac.kr, ryk2001@uos.ac.kr, et2989@uos.ac.kr, eye1719@uos.ac.kr,
*hms@kbs.co.kr

Improving K-POP Fancam Person Detection Performance via Diffusion Model-Based Depth Estimation

Junho Song, Yoonkyung Ra, Beomjun Park, Hanbi Choi, *Minsoo Hong
University of Seoul, *Korean Broadcasting System

요 약

현재 많은 음악 방송에서는 아이돌 무대의 멤버별 직캠 영상을 제작하기 위해 많은 장비와 인력을 소비하고 있다. 이러한 리소스 효율화를 위해 객체 검출 모델을 이용한 직캠 영상 제작 연구가 진행되고 있지만, 무대 위 동선이 겹치는 상황에서는 가려진 객체를 모두 검출하는 데 어려움이 있다. 본 논문에서는 겹친 데이터 중 가장 앞에 있는 데이터를 보존하기 위해 Diffusion 기반의 깊이 추정 모델을 사용하여 2D 이미지에서 깊이 정보를 추출하고, 이를 통해 객체 검출 모델의 성능을 향상시키는 방법을 제안한다. 추가 실험으로 학습 시 데이터 증강 방식으로 Grayscale 을 적용하여 모델의 일반화 성능을 높이는 방법도 제안한다. 제안한 방법을 통해 mAP 를 0.472 에서 0.512 로 향상시켰다.

1. 서론

최근 음악 방송에서는 멤버별 개인 직캠 서비스를 확대하고 있다. 각 아티스트에게 개별 카메라와 촬영 기사를 할당하는 방식으로 직캠 서비스를 제공하고 있는데, 이러한 방식은 많은 인력과 촬영 장비가 필요하며, 무대 공간의 활용에도 제약을 가져올 수 있다.

이에 따라 객체 탐지 기반의 자동 화면 리프레이밍(Re-Framing) 기술을 도입하여 전통적인 촬영 방식을 대체함으로써 비용을 절감하고 무대 공간을 보다 효율적으로 사용하는 방안이 연구되고 있다 [1]. 그러나 무대 영상의 특성상, 안무 도중 발생하는 가려짐(Occlusion)이나 무대 범위를 벗어나는 경우 등이 객체 탐지에 어려움을 야기한다. 또한 세로형 직캠의 특성상 일정한 화면 비율이 유지되어야 하므로 탐지된 객체에 배경의 비율이 높은 경향을 가지고 있다. 그래서 기존에는 객체 검출 모델을 학습할 때, 겹치는 데이터를 제거함으로써 객체의 정보를 누락시키는 방법을 사용했다.

이러한 문제를 해결하기 위해 본 논문은 Diffusion 기반의 깊이 추정 모델을 이용하여 데이터 정제 과정에서 가장 앞에 위치한 인물 데이터를 보존하는 새로운 방법을 제안한다. 이 방법은 인물 간의 깊이를 추정하여 가장 앞의 인물 정보를 보존함으로써, 무대 위 복잡한 동선과 겹침 상황에서 전면의 인물 검출이 가능하도록 한다. 제안된 방법의 유효성은 mAP (mean Average Precision) 지표를 사용하여 기존 모델과의 성능 비교를 통해 향상된 결과를 입증하고 KBS 미디어기술연구소의 VVERTIGO 솔루션 [7]을 이용해 최종 객체 검출 영상을 비교하여 가려짐 문제가 어느정도 해결됨을 입증하고자 한다.

2. 관련 연구

2.1 Object Detection

객체 탐지(Object Detection)는 컴퓨터 비전 기술의 분야 중 하나로 복수의 객체를 분류하고 위치를 파악하는 작업을 수행한다. 객체의 위치를 찾는 것이기 때문에 해당 객체의 경계 상자 (Bounding Box)를 표시하여 위치를 나타낸다. 딥러닝이 화두가 된 이후, CNN 기반 모델이 연구되어 왔고 [2, 3], 최근 Transformer 를 결합한 DETR 기반의 객체 검출 기법들이 등장하였다 [4, 5].

DETR 은 Transformer 의 인코더-디코더 구조를 활용하는 객체 검출 모델로 Facebook AI 에서 2020 년에 제안하였다 [4]. CNN(Convolutional Neural Network) 기반의 모델에서 추출된 이미지 특징이 Transformer 인코더에 입력되어 전역적인 특징을 추출하고 이를 바탕으로 디코더에서 객체를 검출한다.

2.2 Depth Estimation

깊이 추정 (Depth Estimation)은 2D 이미지로부터 3D 깊이 정보를 추정하는 기술로 하나의 이미지로 깊이 정보를 추정하는 Monocular Depth Estimation 과 두 개의 이미지를 이용해 깊이 정보를 추정하는 Stereo Depth Estimation 으로 나뉜다.

최근 Diffusion 모델을 Depth Estimation 분야에 적용하는 연구가 진행되고 있다. Diffusion 모델은 데이터에 잡음을 조금씩 더해가거나 잡음으로부터 조금씩 복원해 가는 과정을 통해 데이터를 생성하는 모델이다.

2023 년 12 월에 공개된 모델 Marigold [6]는 Diffusion 모델을 Depth Estimation 에 적용해 SOTA 를 달성했다. Marigold 는 사전 훈련된 Stable Diffusion 모델을 기반으로 이미지와 깊이 정보를 잠재 공간으로 인코딩하는 VAE (Variational Autoencoder)와 U-Net 을 활용해 깊이 추정을 위한 파인튜닝을 수행한다. 합성 데이터를 학습에 사용하여 정확한 깊이 정보를 생성하고 추론 과정에서는 파인튜닝된 U-Net 을 통해 처리하여 정밀한 깊이 맵을 생성하게 된다 [그림 1]. 또한 이 모델은 제로-샷 학습을 기반으로 학습 데이터에 포함되지 않은 새로운 데이터에서도 높은 수준의 깊이 추정을 수행할 수 있다.

그림 1. Marigold 를 사용한 깊이 추정

3. 데이터

3.1 음악 방송 데이터세트

KBS 뮤직뱅크 무대 이미지 데이터세트를 학습데이터로 사용하며 데이터세트는 무대 영상의 프레임과 각 프레임 내의 객체 좌표 데이터로 구성된다. 기존의 객체 검출 데이터와 달리 세로형 직캠 영상에 맞게 화면비를 유지한 경계 상자의 좌표를 라벨로 사용하므로 경계 상자 내의 배경의 비율이 높은 특징을 가지고 있다. 이에 따라 인물의 이동 동선에 따라 가려짐을 최대한 제거하고 가장 앞쪽 인물의 좌표 데이터를 보존하여 인물 검출 정확도를 높이고자 한다.

3.2 깊이 추정 기반 데이터 정제

데이터 정제의 순서는 [그림 2]와 같다. 우선 데이터 내에 경계 상자 좌표 내부에 사람이 없어 학습 데이터로 사용하기 부적절한 데이터들이 존재한다. 경계 상자 내부에 사람이 없는 경우 경계 상자를 제거하여 초기 정제 과정을 거친다.

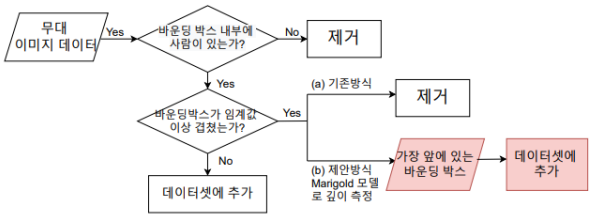

그림 2. 데이터 정제 순서도

정제 과정 후, 데이터세트 정제 방식에 따른 객체 검출 결과를 확인하기 위해 경계 상자들이 겹쳐 있는 경우, IoU (Intersection over Union) 기준 임계 값 이상 겹치는 경계 상자는 Marigold 모델을 사용하여 깊이를 측정하고 가장 앞에 있는 경계 상자를 보존하여 정제된 데이터세트를 생성한다.

원본 이미지를 Marigold 모델에 통과시키면 깊이 정보가 포함된 픽셀 값이 생성된다 [그림 3]. 픽셀 값을 통해 각 객체 간 상대적 위치를 판단할 수 있으며 겹친 경계 상자들을 분리하기 위한 기준으로 사용한다. 겹친 경계 상자 중에서 평균 픽셀 값이 가장 작은 경계 상자를 가장 앞에 있는 객체로 판단하여 추출한다. 그 결과, IOU 가 특정 임계값 이상 겹치는 box 를 모두 제거하는 기존 방식과는 달리 겹쳐진 객체들을 모두 제거하지 않고 가장 앞에 있는 객체를 추출할 수 있다 [그림 4].

그림 3. 원본 이미지(상)와 Marigold 적용한 이미지(하)

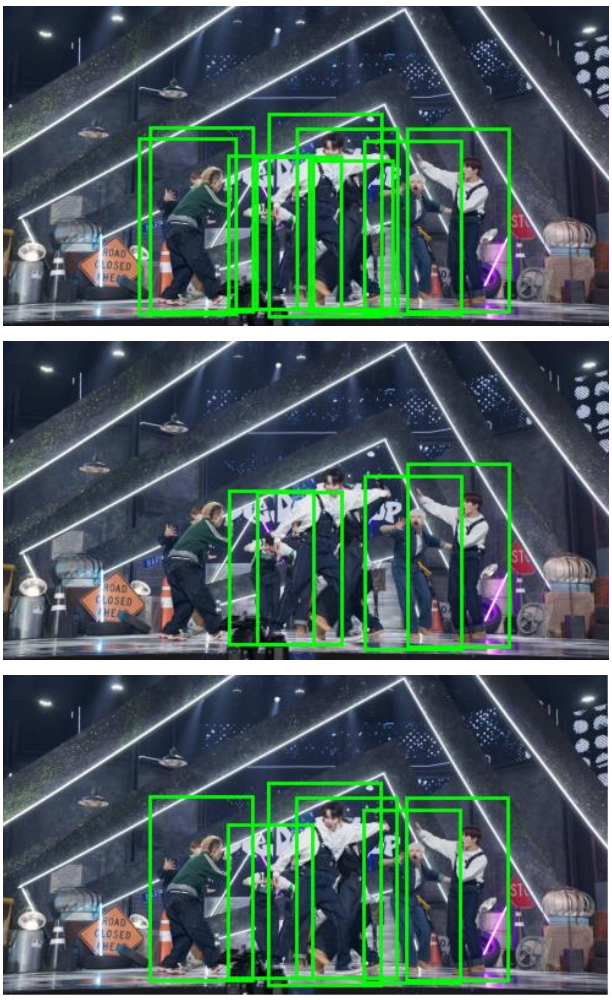

그림 4. 원본 데이터 경계 상자(상), 기존 방식의 경계 상자 (중), 깊이 추정을 적용한 경계 상자 (하)

4. 실험

4.1 실험 환경

인물 검출을 위한 모델로 Co-DETR [5]을 선정하였고 데이터 세트의 겹친 경계 상자 정제를 위해 Diffusion 기반 깊이 추정 기법 Marigold 를 사용한다.

두 모델의 학습 및 추론을 위해 리눅스 기반의 NVIDIA RTX A5000 을 사용하였다. 최종적으로 실험 결과 검증 및 시각화를 위하여 KBS 의 VVERTIGO 솔루션[7]을 이용해 최종 객체 검출 영상을 도출한다.

4.2 실험 설계

Diffusion 기반 깊이 추정 기법을 통한 데이터세트 정제의 효과를 확인하기 위해 객체 검출 모델, 학습 파라미터 및 데이터 세트 분할 비율은 고정하여 실험을 진행한다.

학습 파라미터는 lr 는 2e-05, 배치 사이즈 2, 스케쥴러로 MultistepLR 을 사용하고 총 12 epochs 동안 학습을 진행하며 데이터세트는 총 세 가지로 나누어 모델을 학습한다.

1. 기존의 겹친 데이터가 전부 제거된 비정제 데이터
2. Diffusion 기반 깊이 추정이 적용된 정제 데이터
3. Diffusion + Grayscale 기반 정제 데이터

훈련, 검증, 테스트 세트 비율을 8:1:1 로 설정하며 비슷한 영상데이터의 편향을 막고 고른 분포를 가질 수 있도록 무대 별 층화추출하여 학습 시 중복 데이터가 최소화되도록 구성하였다. 최종 성능 평가 때는 학습에 사용되지 않은 새로운 뮤직뱅크 무대 데이터 1,000 장을 사용하여 성능을 비교한다.

4.3 실험 결과

객체 탐지 분야에서 일반적으로 사용하는 mAP 를 평가 지표로 사용한다. 세분화된 성능 확인을 위해 IoU 값에 따라 mAP_50, mAP_75 로 나눈다.

	epochs	mAP	mAP_50	mAP_75
baseline	12	0.472	0.689	0.581
diffusion	12	**0.512**	**0.744**	**0.621**
diffusion + grayscale	12	**0.519**	**0.760**	**0.630**

표 1. 실험 결과

비정제 데이터세트와 비교했을 때 본 논문이 제안한 Diffusion 기반 정제 데이터로 학습한 경우, mAP 에서 더 높은 성능을 보이는 것을 확인할 수 있다 [표 1]. 이는 비정제 데이터세트를 사용한 경우 겹치는 객체에 대해 정보 손실이 발생했다는 것을 의미하고 겹치는 객체의 적합한 경계 상자를 찾는 것이 모델 향상에 영향을 미친다는 것을 의미한다.

추가로, Grayscale 을 적용한 데이터 증강 방식을 통해 다양한 조명 및 무대 환경에서의 객체 검출 모델의 일반화 성능을 끌어올렸다. 색상 정보의 감소로 인해 조명, 경계 등의 특징이 강조되어 학습되기 때문에 다양한 배경 색상이나 조명 조건에서도 강인하게 대처할 수 있다.

그림 5. 기존 방식(좌)과 제안한 방식(우) 비교

기존 방식에서는 주 검출 인물이 겹친 경계 상자의 영향으로 중심에서 벗어나 있지만, 제안한 Diffusion 모델을 적용한 방식에서는 주 검출 인물이 최대한 중심에 있도록 검출한다 [그림 5]. 결과적으로, VVERTIGO 를 통해 직캠 영상을 생성할 때, 비정제 데이터세트를 사용한 모델과 비교하여 정제 데이터세트를 사용한 모델은 겹친 상황에서 주 검출 인물을 직캠의 중앙에 위치하게 하여 시각적 집중도를 높일 수 있다.

5. 결론

본 논문에서는 Diffusion 모델을 기반으로 깊이 추정을 통해 2D 이미지 데이터세트를 정제하고, 이를 이용해 학습한 객체 검출 모델의 검출 정확도를 향상시켰다. 이 방법을 통해 직캠 AI 자동화 성능을 개선시키는 방법을 제안하였다. 깊이 추정 모델을 활용하여 겹치는 경계 상자 중 가장 앞의 상자를 보존하여 무대 영상에서 인물들의 동선이 겹쳤을 때 인물이 검출되지 않는 문제를 개선했다.

실험 결과, 제안한 방법은 기존의 비정제 데이터를 사용한 방식에 비해 mAP 에서 0.472 에서 0.512 까지 성능 향상을 보였으며 Grayscale 을 포함한 데이터 증강 기법을 통해 모델의 일반화 능력을 향상시켜 다양한 조명과 배경 조건에서도 강인한 성능을 낼 수 있도록 하였다.

6. Acknowledgement

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 디지털 신기술 인재양성 혁신공유대학사업 (차세대통신)의 연구 결과입니다.

7. 참고 문헌

[1] "AI 기반 멀티뷰 제작시스템 VVERTIGO 를 소개합니다." 방송과 기술. 2022 년 10 월호.

[2] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[3] Tan, Mingxing, Ruoming Pang, and Quoc V. Le. "Efficientdet: Scalable and efficient object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[4] Carion, Nicolas, et al. "End-to-end object detection with transformers." *European conference on computer vision*. Cham: Springer International Publishing, 2020.

[5] Zong, Zhuofan, Guanglu Song, and Yu Liu. "Detrs with collaborative hybrid assignments training" *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.

[6] Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation." *arXiv preprint arXiv:2312.02145* (2023).

[7] https://www.vvertigo.com