JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 1, pp. 77-89, January 2026
https://doi.org/10.9708/jksci.2026.31.01.077

A Real-time Head Pose Estimation via YOLO-based Facial Landmark Detection using High-Fidelity Synthetic Data

Un-Yong Kim*,***, Sungkuk Chun**, Jeongrok Yun*, Ju-Yeong Park*, Sung-Hoon Hong****, Hoe-Min Kim*****

*Researcher, Spatial Optical Information Research Center, Korea Photonics Technology Institute, Gwangju, Korea
**Senior Researcher, Spatial Optical Information Research Center, Korea Photonics Technology Institute, Gwangju, Korea
***Master's Candidate, Department of Electronics and Computer Eng, Chonnam National University, Gwangju, Korea
****Professor, Department of Electronics and Computer Eng, Chonnam National University, Gwangju, Korea
*****Principal Researcher, Spatial Optical Information Research Center, Korea Photonics Technology Institute, Gwangju, Korea

[Abstract]

Data collection constraints and the Sim2Real gap are primary challenges in developing head pose estimation systems. This study adopts geometric landmark information, which is robust against visual noise, as a core feature to bridge this gap. The methodology consists of a two-stage pipeline. First, a large-scale synthetic dataset is constructed using Unreal Engine and MetaHuman. Second, the YOLOv11-pose model is trained as a facial landmark detector using a mixture of synthetic data and the real-world BIWI dataset. The system then estimates the three-axis angles—Roll, Pitch, and Yaw—in real-time based on the detected landmark coordinates. In evaluations using the BIWI dataset, the model achieved a low Mean Absolute Error (MAE) of 1.00° in the near-frontal region. Furthermore, the final system ensured a real-time processing speed of 21.2 FPS in a webcam environment. In conclusion, the integration of synthetic and real data with a landmark-based approach demonstrates the feasibility of precise, real-time head pose estimation.

▸ **Key words**: Head Pose Estimation, Synthetic Data, YOLO, Landmark, Unreal Engine, MetaHuman

• First Author: Un-Yong Kim, Corresponding Author: Sung-Hoon Hong, Hoe-Min Kim
*,***Un-Yong Kim (kuy7023@kopti.re.kr), Spatial Optical Information Research Center, Korea Photonics Technology Institute, Department of Electronics and Computer Eng, Chonnam National University
**Sungkuk Chun (k612051@kopti.re.kr), Spatial Optical Information Research Center, Korea Photonics Technology Institute
*Jeongrok Yun (justin182@kopti.re.kr), Spatial Optical Information Research Center, Korea Photonics Technology Institute
*Ju-Yeong Park (zhseltm@kopti.re.kr), Spatial Optical Information Research Center, Korea Photonics Technology Institute
****Sung-Hoon Hong (hsh@jnu.ac.kr), Department of Electronics and Computer Eng, Chonnam National University
*****Hoe-Min Kim (hmkim@kopti.re.kr), Spatial Optical Information Research Center, Korea Photonics Technology Institute
• Received: 2025. 11. 11, Revised: 2025. 12. 01, Accepted: 2025. 12. 31.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

데이터 수집 제약과 Sim2Real 간극은 안면 방향 추정 시스템 개발의 주요 과제이다. 본 연구는 이 간극을 해소하고자 시각적 노이즈에 강건한 기하학적 랜드마크 정보를 핵심 특징으로 채택하였다. 방법론은 2단계 파이프라인으로 구성된다. 첫째, Unreal Engine과 MetaHuman을 활용하여 대규모 합성 데이터셋을 구축한다. 둘째, 합성 데이터와 실제 BIWI 데이터를 혼합하여 YOLOv11-pose 모델을 얼굴 랜드마크 검출기로 학습시키고, 랜드마크 좌표 기반으로 3축 각도 Roll, Pitch, Yaw를 실시간 추정한다. 실제 BIWI 데이터셋 검증에서 안면방향추정 모델은 Near-frontal 영역을 대상으로 한 검증에서 1.00도의 낮은 MAE를 보였으며, 최종 시스템은 웹캠 환경에서 21.2 FPS의 실시간 처리 속도를 보였다. 결론적으로, 합성 및 실제 데이터 결합과 랜드마크 기반은 정밀한 안면 방향 추정 기술의 실시간 적용 가능성을 제시한다.

▸ **주제어**: 안면방향추정, 합성데이터, YOLO, 랜드마크, Unreal Engine, MetaHuman

I. Introduction

최근 인공지능 분야는 모델의 구조적 발전을 넘어 데이터의 질과 양이 모델의 성능을 결정하는 데이터 중심 패러다임으로 전환되고 있다[1]. 특히 인간의 시각적 정보를 이해하는 컴퓨터 비전 기술에서 이러한 경향은 더욱 두드러진다. 수많은 컴퓨터 비전 응용 분야 중에서도, 사용자의 시선과 주의 집중 영역을 파악하는 안면 방향 추정(Head Pose Estimation) 기술은 인간과 기계의 직관적인 상호작용을 위한 핵심 요소로 자리 잡았다. 이 기술은 운전자의 상태를 실시간으로 모니터링하는 지능형 차량 시스템부터, 사용자의 시야에 가상 정보를 정밀하게 증강시키는 AR/VR 환경에 이르기까지 그 중요성이 날로 증가하고 있다.

그러나 안면 방향 추정 기술의 고도화는 훈련 데이터 확보라는 근본적인 장벽에 부딪히고 있다. 딥러닝 모델이 높은 일반화 성능을 갖추기 위해서는 다양한 인종, 연령, 표정과 더불어 예측 불가능한 조명 및 가려짐 상황을 포함하는 방대한 데이터셋이 필수적이다[3][4]. 현실 세계에서 이러한 데이터를 직접 수집하는 것은 막대한 시간과 비용을 소모할 뿐만 아니라, 개인의 얼굴이라는 민감한 생체 정보를 다루기에 엄격한 개인정보보호 규제와 윤리적 문제를 동반한다. 설령 데이터 수집에 성공하더라도, 이미지 속 인물의 머리 방향(Roll, Pitch, Yaw)을 정확한 3차원 각도 값으로 라벨링하는 과정에서 발생하는 측정 오차는 데이터의 신뢰성을 저해하는 고질적인 문제로 남아있다[5].

이러한 현실 데이터의 한계를 극복하기 위한 대안으로, 정밀한 제어가 가능한 가상 환경에서 생성된 합성 데이터(Synthetic Data)가 새로운 가능성을 제시한다[6]. 언리얼 엔진[7]과 같은 현대적 3D 렌더링 기술은 현실과 유사한 수준의 고품질 그래픽을 제공하며, 이는 데이터 생성의 패러다임을 바꾸고 있다. 합성 데이터의 가장 큰 강점은 데이터의 구성 요소를 수학적으로 정밀하게 제어할 수 있다는 점이다. 가상 환경에서는 조명, 카메라 각도, 배경 등 외부 환경은 물론, 가상 인물의 머리 방향 값을 오차 발생 가능성이 매우 낮은 상태로 추출하여 높은 정밀도의 Ground Truth를 확보할 수 있다. 이는 데이터 자체의 모호함이나 오류를 최소화하여, 모델이 문제의 본질에 더 집중하여 학습할 수 있는 이상적인 환경을 제공한다.

물론, 합성 데이터로 학습한 모델을 실제 환경에 적용할 때 발생하는 현실과 가상의 간극(Sim2Real Gap)은 여전히 중요한 학술적 과제이다[8]. 본 논문에서는 이 간극을 최소화하기 위한 전략으로, 이미지의 질감이나 색감과 같은 시각적 요소보다 변화에 강인한 기하학적 특징, 즉 얼굴 랜드마크에 집중하는 접근법을 제안한다. 우리는 언리얼 엔진을 통해 생성된 3D 아바타로부터 정밀한 랜드마크 좌표와 그에 상응하는 3차원 머리 방향 값을 추출하여 데이터셋을 구축하였다. 이후, 이 데이터셋을 통해 YOLOv11-pose 프레임워크 기반 얼굴 랜드마크 추론 모델을 학습하였다. 최종적으로, 학습된 모델을 안면 방향 추정 모델과 통합하여, 실시간 영상 내 사용자의 안면 방향을 추정하고 가시화함으로써 제안하는 합성 데이터 활용 방법론의 실용성과 효용성을 검증하고자 한다.

본 논문의 구성은 다음과 같다. 2장 관련 연구에서는 딥러닝 기반 안면 방향 추정 기술의 흐름과 YOLO 기반 랜드마크 검출 동향을 확인하고, 합성 데이터 생성 및 Sim2Real Gap 완화를 위한 기존 연구 사례를 분석한다.

3장 제안 방법에서는 Unreal Engine과 MetaHuman을 활용한 고품질 합성 데이터 생성 파이프라인과 YOLOv11-pose 기반의 랜드마크 검출 및 DNN 기반 안면 방향 추정 모델의 통합 구조를 기술한다. 4장 실험 및 결과에서는 데이터 전처리 과정과 실제 데이터 및 합성 데이터의 최적 혼합 비율 도출 실험을 설명하고, 기존 연구 모델들과의 정량적 성능 비교 및 실시간 처리 속도 측정 결과를 제시한다. 5장 결론에서는 본 연구의 주요 성과를 요약하고 향후 연구 과제를 제시한다.

II. Related Works

본 장에서는 제안하는 방법론의 이론적 배경을 입증하기 위해, 안면 방향 추정 분야의 주요 선행 연구들을 네가지 관점에서 분석한다. 첫째, 딥러닝 기반 안면 방향 추정 연구의 기술적 흐름을 살펴본다. 둘째, 랜드마크 추출을 위한 YOLO 기반 연구 동향을 확인한다. 셋째, 데이터셋 구축 방식과 각 방식의 한계를 고찰한다. 마지막으로, 합성 데이터의 고질적인 문제인 Sim2Real Gap을 해결하기 위한 기존 연구들을 설명한다.

2.1. Deep Learning-based Head Pose Estimation

기존 안면 방향 추정 연구는 크게 정확도 중심의 접근법과 경량화 및 실시간 처리를 위한 접근법으로 나뉜다.

2.1.1 Accuracy-focused Approaches

HopeNet은 안면 방향 추정을 연속적인 회귀 문제와 분류 문제로 함께 푸는 방식을 제안했다[9]. 최종 각도는 각 구간에 속할 확률의 기댓값으로 계산하고, Roll, Pitch, Yaw 각각에 대한 다중 손실 함수를 도입하여 안정적인 학습을 유도했다. 이 방식은 큰 각도에서도 비교적 안정적인 성능을 보이지만, 각도 구간의 경계에서 오차가 커질 수 있고 양자화로 인해 매우 정밀한 각도 예측에는 한계가 있다.

FSA-Net은 특징 맵을 여러 그룹으로 나누어 각 그룹이 특정 각도 범위의 특징을 집중적으로 학습하도록 하는 Fine-Grained Structure Aggregation 모듈을 제안하여 큰 각도에서의 성능을 향상시켰다[10]. 별도의 복잡한 구조 없이 특징 맵을 효율적으로 활용하지만, 여전히 단일 이미지의 외형적 특징에만 의존하므로 가려짐이나 비정상적인 조명에는 취약할 수 있다.

TriNet은 오일러 각도 대신 회전 행렬 표현을 직접 추정하는 랜드마크-프리 접근 방식을 제안했다. 이들의 모델은 회전 행렬을 구성하는 3개의 단위 벡터를 예측하도록 학습하며, 예측의 안정성을 높이기 위해 추가적인 직교성 손실을 통합했다[11].

6DRepNet은 오일러 각도의 모호성 문제를 해결하기 위해 연속적인 6D 회전 행렬 표현을 직접 회귀하는 방식을 제안하여 SOTA 수준의 정확도를 보였다[12]. 하지만 모델이 360° 전체 범위 예측이 가능하도록 설계되었지만, 학습에 사용된 300W-LP와 같은 표준 데이터셋이 주로 정면 뷰에 편향되어 있습니다. 이로 인해 모델의 전체 범위 예측 잠재력을 최대로 활용하여 학습하고 검증하는 데 한계가 있다[12].

2.1.2 Lightweight and Real-time Approaches

MobileNets는 표준 컨볼루션을 깊이별 및 점별 컨볼루션으로 분리하는 방법을 제안하여, 연산량을 획기적으로 줄이면서도 성능 저하를 최소화한 경량 네트워크 아키텍처이다[13]. 모바일 및 임베디드 기기에서 실시간으로 동작할 수 있을 만큼 매우 효율적이나, 경량화를 위해 표현력이 감소하여 복잡한 모델에 비해 정확도가 낮다는 근본적인 한계를 가진다.

ShuffleNet은 점별 컨볼루션을 그룹 컨볼루션으로 바꾸고, 채널 셔플링 연산을 도입하여 채널 간 정보 교환을 촉진함으로써 MobileNet보다 더 높은 효율성과 정확도를 달성했다[14].

LwPosr은 안면 방향 추정 문제에 특화된 경량 CNN 아키텍처를 제안하여, 빠른 추론 속도를 유지하면서도 준수한 수준의 정확도를 달성하는 데 집중했다[15].

2.2. YOLO-based Landmark Detection

YOLO(You Only Look Once)는 본래 실시간 객체 탐지(Object Detection)를 위해 개발된 1-stage 딥러닝 모델이다[16]. 초기 YOLO 버전들은 바운딩 박스 예측에 중점을 두었으나, 이후 YOLOv5-Face와 같은 일부 연구에서는 YOLO의 출력 레이어를 수정하여 바운딩 박스와 랜드마크 좌표를 동시에 예측하도록 개조하였다[17].

최근에는 YOLOv7-pose[18], YOLOv8-pose[19], 그리고 본 연구에서 활용한 YOLOv11-pose[20]와 같이, 객체 탐지와 키포인트 추정 작업을 하나의 통합된 아키텍처로 처리하는 포즈 추정(Pose Estimation) 모델로 공식적으로 확장되었다. 이러한 모델들은 1) 얼굴 탐지 모델과 2) 랜드마크 추출 모델을 별도로 실행하는 2-stage 접근법에 비해, 단일 모델이 두 가지 작업을 한 번에 처리하므로 추론 속도가 매우 빠르다는 장점이 있다. 본 연구는 이러한

YOLO-pose 모델의 빠른 추론 속도와 높은 정확도를 활용하여, 웹캠 환경에서도 안정적으로 23개의 주요 랜드마크를 추출하는 검출기로 사용하였다.

2.3. Data Generation

2.3.1 Real-world Datasets

BIWI Kinect Head Pose Database는 Microsoft Kinect 센서를 사용하여 20명의 피실험자로부터 수집한 RGB-D 데이터셋이다[21]. 깊이 정보를 활용하여 비교적 정확한 3D 라벨을 제공하지만, 통제된 실내 환경에서 소수의 인원만을 대상으로 하여 데이터 다양성이 부족하다는 한계가 있다.

AFLW2000은 실제 인터넷 환경의 이미지 2,000개에 3DMM을 이용해 랜드마크와 3축 각도를 라벨링하였다[5]. 다양한 실제 환경을 포함하지만, 3DMM으로 생성된 라벨의 정확도 한계와 딥러닝 학습에 부족한 데이터 규모가 단점이다.

300W는 안면 랜드마크 검출 연구의 표준 벤치마크로, 다양한 실제 환경을 포함하지만, 규모가 작아 주로 평가용으로 활용된다[22].

2.3.2 Synthetic Datasets

UnityEyes는 Unity 엔진을 활용하여 시선 추정을 위한 대규모 합성 눈 이미지를 생성했으나, 얼굴 전체가 아닌 눈 영역에 국한되며 당시 렌더링 기술의 한계로 현실성이 다소 부족했다[23]. 3DMM-based Generation 방식은 평균 얼굴 모델을 기반으로 하여 개인의 고유한 특징이나 머리카락 등 비선형적 요소를 세밀하게 표현하는 데 한계가 있었다[24]. StyleGAN은 매우 사실적인 2D 얼굴 이미지를 생성할 수 있지만, 동일 인물의 3D 일관성을 유지하며 각도를 정밀하게 제어하기 어렵다는 근본적인 한계를 가진다[25]. HeadNeRF는 새로운 시점의 이미지를 사실적으로 렌더링할 수 있지만, 특정 개인을 모델링하는 데 최적화되어 있어 대규모의 다양한 데이터셋을 구축하기에는 어려움이 있다[26].

2.4. Bridging the Sim2Real

2.4.1. Domain Adaptation

Sim2Real Gap을 해결하는 주요 접근법 중 하나는 도메인 적응으로, 라벨이 없는 실제 데이터에 모델을 적응시키는 기법이다. DANN은 특징 추출기가 도메인에 무관한 특징을 학습하도록 적대적 학습 방식을 제안했다[27]. CycleGAN은 쌍을 이루지 않은 이미지 간의 변환을 학습하여, 합성 이미지를 실제 이미지의 스타일로 변환하는 방식을 사용했다[28]. ADDA는 소스와 타겟 도메인의 매핑 함수를 분리하여 더 유연한 적응을 시도했다[29]. 그러나 이러한 연구들은 복잡한 모델 구조와 불안정한 적대적 학습에 의존하며, 기하학적 정보가 왜곡될 수 있는 한계가 있다.

2.4.2. Hybrid Data Learning Approach

다른 Sim2Real Gap 해결책은 합성 데이터와 실제 데이터를 함께 사용하여 학습하는 혼합 데이터 학습 방식이다.

Playing for Data는 컴퓨터 게임(GTA-V)에서 대규모 라벨링 데이터를 추출하여 실제 데이터와 함께 사용함으로써, 합성 데이터의 가능성을 보여준 선구적인 연구이다[8]. SimGAN은 적대적 학습을 통해 합성 이미지의 현실성을 높이는 정제기 네트워크를 제안했다[30]. Domain Randomization은 시뮬레이션 환경의 조명, 질감 등을 매우 넓은 범위에서 무작위로 변화시켜, 모델이 현실 세계를 시뮬레이션의 또 다른 변형 중 하나로 인식하게 만드는 전략이다[31].

본 연구는 이러한 접근법들, 특히 포토리얼리스틱 렌더링을 통해 도메인 차이 자체를 최소화하고, 소량의 실제 데이터를 직접 학습에 포함하는 혼합 데이터 방식을 사용하여 더 직관적이고 안정적인 Sim2Real Gap 해결책을 제시한다.

III. The Proposed Method

본 연구는 실제 데이터 수집 과정에서 발생하는 윤리적, 비용적 문제를 해결하고, 강건한 안면 방향 추정 시스템을 구축하기 위해 합성 데이터 기반의 딥러닝 파이프라인을 새롭게 제안한다. 제안하는 시스템은 고품질 합성 데이터 생성 파이프라인, YOLOv11 기반 얼굴 랜드마크 검출 모델 학습 및 추론, 그리고 3차원 안면 방향 추정을 위한 모델 학습 및 추론 3단계로 구성된다. Fig. 1.은 전체 파이프라인의 구성도이다.

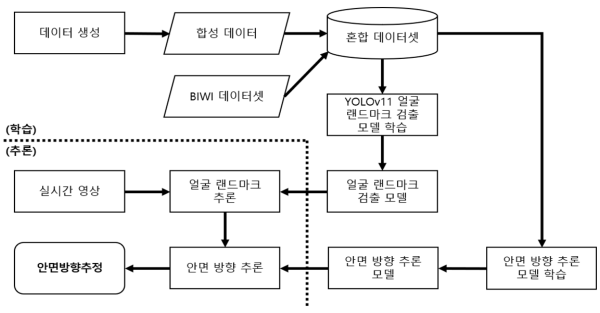

Fig. 1. Pipeline of the Proposed Real-time Head Pose Estimation

3.1. Data Generation

정확한 딥러닝 모델 학습의 전제 조건은 데이터의 품질과 다양성입니다. 본 연구에서는 기존 합성 데이터 생성 기술들의 현실성 및 제어 가능성 한계를 극복하고자, 합성 데이터와 실제 데이터셋인 BIWI를 혼합하여 데이터셋을 구성하였다.

3.1.1 Virtual Environment and 3D Model Selection

데이터 생성 환경으로는 영화 및 게임 산업에서 최고 수준의 시각적 결과물을 위해 사용되는 언리얼 엔진(Unreal Engine)을 채택하였다. 언리얼 엔진은 재질의 물리적 특성을 시뮬레이션하는 물리 기반 렌더링(PBR)과 빛의 반사와 굴절을 실시간으로 추적하는 루멘 글로벌 일루미네이션 시스템을 통해 언리얼 데모의 한 장면인 Fig. 2.와 같은 퀄리티를 보여주어 배경을 제작했을 때 실제 사진과 거의 구분이 어려운 수준의 결과물을 보장한다.

Fig. 2. Unreal demo Image

데이터의 주체가 될 가상 인간 모델로는 언리얼 엔진에서 제공하는 메타휴먼 크리에이터(Metahuman Creator)를 활용하였다. 메타휴먼 크리에이터는 실제 인간의 3D 스캔 데이터를 기반으로 제작되었다. 캐릭터 생성 장면인 Fig. 3.의 이미지에서 보여주는 바와 같이 캐릭터를 생성할 때 피부의 미세한 질감부터 다양한 머리 스타일과 눈과 눈썹 모양 및 정교한 표정까지 표현할 수 있으며, 이는 Sim2Real Gap을 최소화하는 핵심적인 요소로 작용한다.

Fig. 3. MetaHuman demo Image

3.1.2 Automated Data Generation Pipeline

데이터의 대량 생성을 위해 언리얼 엔진의 블루프린트(Blueprint) 비주얼 스크립팅 시스템을 활용하여 전체 생성 과정을 자동화하였다[32]. 각 축(Roll, Pitch, Yaw)에 대해 -21°부터 +21°까지 3° 간격의 15개 기준 각도를 설정하고, 모델이 특정 값에 과적합 되는 것을 방지하고자 각 구간 내에서 무작위 샘플링을 적용하여 총 3,375개의 고유한 각도 조합을 생성하였다. 이 과정은 아시아, 아프리카, 유럽 등 각기 다른 인종, 연령, 성별의 특성을 가진 메타휴먼 캐릭터 12종에 대해 반복 수행하여, 데이터 편향을 최소화한 총 40,500장의 이미지 데이터셋을 구축하였다. 이 파이프라인의 핵심은 렌더링 된 이미지와 함께, 해당 시점의 오차가 없는 정밀한 랜드마크 좌표 및 3축 각도 값(Ground Truth)이 동시에 저장된다는 점입니다. Fig. 4.는 이 파이프라인의 블루프린트 구조 이미지이다.

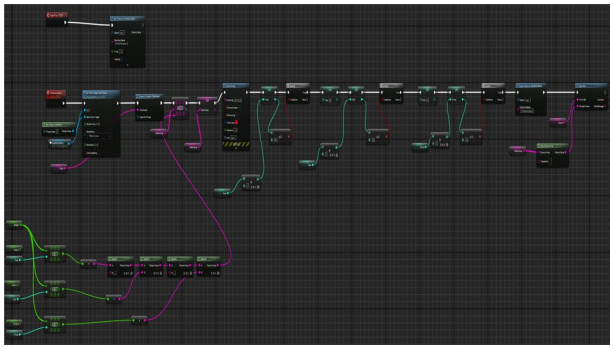

Fig. 4. Blueprint structure

3.1.3 Mixed Dataset Configuration

Sim2Real Gap을 완화하고 실제 얼굴에 대한 적응력을 높이고자, 1단계에서 생성한 40,500장의 합성 데이터셋과 함께 실제 데이터셋인 BIWI의 일부를 학습에 사용하였다. Fig. 5.의 왼쪽은 합성 데이터셋 오른쪽은 BIWI 데이터셋의 샘플 이미지다. BIWI 데이터셋 중, 합성 데이터와 각도 분포를 맞추기 위해 Roll, Pitch, Yaw 값이 모두 -21°에서 +21° 사이인 2,715개의 이미지를 선별하여 훈련 데이터에 추가하였다. 이 방식은 대규모 합성 데이터를 통해 다양한 각도와 인물에 대한 일반화 성능을 확보하는 동시에, 소량의 실제 데이터를 통해 모델이 실제 카메라 센서의 노이즈나 질감 차이와 같은 현실 세계의 특성에 적응하도록 하였다.

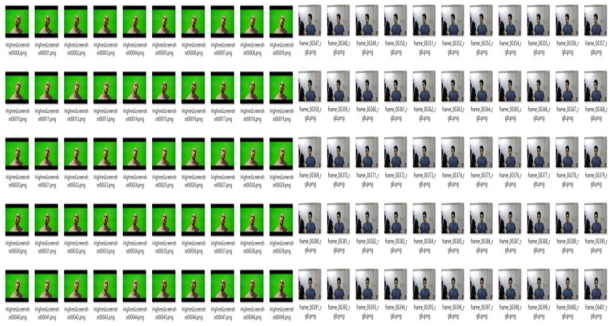

Fig. 5. Synthetic and BIWI Image Samples

3.2. YOLOv11-based Facial Landmark Detection

본 시스템의 핵심 구성 요소는 이미지로부터 안면 방향 추정에 필요한 기하학적 특징을 신속하고 정확하게 추출하는 랜드마크 검출기다. 이를 위해 최신 객체 탐지 모델인 YOLOv11-pose 모델을 랜드마크 검출 문제에 맞게 적용하고, 앞의 3.1.3에서 생성한 합성 데이터와 실제 데이터를 결합한 혼합 데이터셋을 이용해 학습을 진행한다. 혼합 데이터셋을 사용함으로써 Sim2Real Gap을 완화하였고 안면 방향 추정을 위한 랜드마크 검출을 진행하였다.

3.2.1 Model Training

YOLOv11-pose 모델은 이 혼합 데이터셋을 기반으로, 안면 방향 추정에 필요한 23개의 주요 랜드마크 위치를 정확하게 예측하도록 훈련되었다. 모델 학습을 위한 라벨은 각 데이터셋에서 추출된 랜드마크 좌표를 정규화(Normalization)하여 생성하였다. 이를 통해 모델은 이미지의 크기나 해상도에 상관없이 일관된 스케일의 좌표를 학습하게 되며, 이는 실제 환경에서의 강건성을 높이는 데 기여한다.

3.2.2 Facial Landmark Detection

학습이 완료된 YOLOv11-pose 모델은 실시간 웹캠 비디오 스트림에서 각 프레임을 입력받아 안면 영역과 23개의 주요 랜드마크를 동시에 검출한다.

검출 과정에서 모델은 얼굴 영역을 bounding box로 식별하고, 각 랜드마크의 픽셀 좌표를 출력하였다. 이 랜드마크 좌표는 후속 DNN 모델의 입력으로 사용되기 전에 3.2.2절에서 언급된 학습 과정과 동일하게 정규화(Normalization) 과정을 거쳤다. 이는 이미지 해상도나 얼굴 크기의 변화에 따른 좌표 값의 변동을 [-1, 1] 범위로 통일시켜, DNN 추정 모델의 강건성과 일관성을 확보하는 데 필수적이었다.

최종적으로, 검출 및 정규화된 23개의 랜드마크 좌표는 2차원 공간의 특징 벡터로 결합되어 총 46차원의 입력 벡터 형태로 안면 방향 추정 모델에 전달된다. 벡터는 얼굴의 기하학적 자세 정보만을 담고 있어, 픽셀 단위의 Sim2Real Gap 영향을 최소화하며 높은 정밀도의 자세 추정을 가능하게 하였다.

3.3. Head Pose Estimation

제안하는 시스템은 학습된 얼굴 랜드마크 검출기와 기존 연구를 통해 확보된 안면 방향 추정 모델을 결합하여 실시간 영상으로 안면 방향 추정 하는 통합 파이프라인이다. 이 파이프라인은 웹캠으로부터 입력을 받아 최종적인 3축 각도 값을 시각화하기까지 세 단계를 거친다.

제안 된 방법은 첫 번째로 웹캠에서 프레임을 얻어와 학습된 얼굴 랜드마크 검출 모델에 입력된다. 이를 통해 입력된 이미지 내에서 사용자의 얼굴을 탐지하고 안면 방향 추정에 필요한 23개의 주요 랜드마크 좌표를 추출한다.

마지막으로 정규화 과정을 거쳐 생성된 46차원 랜드마크 벡터(23개 랜드마크 × 2차원 좌표)는 DNN 기반의 안면 방향 추정 모델에 최종 입력으로 전달된다. 이 모델은 입력된 랜드마크 벡터를 분석하여 3축 각도(Roll, Pitch, Yaw) 값을 추정한다.

이를 위한 안면 방향 추정 모델은 사전에 학습된 DNN 모델을 기반으로 제안된 방법을 통해 구축한 데이터베이스로 재학습 및 최적화를 진행하였다[33]. 이 모델은 본 연구에서 구축한 3축(Roll, Pitch, Yaw) 각도가 -21도에서 +21도 범위 내에 이미지로 이루어진 합성 데이터셋과 실제 데이터셋(BIWI)을 포함하는 통합 데이터셋에서 추출된 랜드마크를 이용하여 학습되었으며, 학습 모델은 DNN 모델을 이용해 학습을 진행했다. 기존 제안 방법의 입력은 18개의 랜드마크를 입력으로 했지만 본 연구에서는 모델의 성능을 향상 및 최적화를 위해 앞선 YOLOv11 모델 학습에 사용한 23개의 랜드마크 좌표를 입력으로 학습을 진행했고 과적합 방지를 위해 L2정규화, Batch Normalization 및 Dropout을 적용했다.

IV. Experiments

본 연구의 실험은 최종 안면 방향 추정 시스템의 구축 과정을 따라 합성 데이터 생성부터 3축 각도 예측까지의 전체 파이프라인을 검증하는 구조로 진행되었다.

1. Dataset Configuration and Preprocessing

랜드마크 추출 모델 학습을 위해 두 가지 종류의 데이터셋을 조합하여 사용하였다. 첫 번째 데이터셋은 자체 제작

한 3D 아바타를 Unreal Engine 환경에서 렌더링하여 생성한 합성 이미지 데이터이다. 총 12개의 캐릭터 데이터를 사용하였다. 각 캐릭터는 약 3,375개의 이미지 프레임을 포함한다. 두 번째 데이터셋은 BIWI 데이터셋이다. 실제 사람의 얼굴 이미지와 해당 이미지에서의 머리 방향 (Roll, Pitch, Yaw) 각도 값을 포함하는 공개 데이터셋이다. 이와 같은 두 개의 데이터셋에 대하여 학습을 위한 데이터전처리를 진행하였다.

데이터 전처리 과정은 다음과 같다.

- 각도 필터링(BIWI)
  BIWI 데이터셋의 경우, 각 인물 폴더에 대응하는 각 이미지 프레임별 머리 방향각도 정보를 담은 파일을 참조하여 Roll, Pitch, Yaw 값이 모두 -21도에서 +21도 범위 내에 있는 이미지 프레임만 선별하여 사용하였다. 이는 극단적인 각도에서 발생할 수 있는 랜드마크 추출 오류 및 가려짐 현상을 최소화하기 위함이다.
- 랜드마크 추출
  선별된 모든 이미지(Unreal 및 필터링된 BIWI)에 대해 얼굴 랜드마크를 추출하였다. 다양한 각도(Pitch, Yaw)에서의 안정적인 추출을 고려하여, 최종적으로 코, 눈, 입술, 눈썹 부위의 23개 주요 랜드마크를 선택하여 사용하였다 Fig. 6.은 Unreal 이미지 샘플에 랜드마크를 가시화한 샘플 이미지이고 각 랜드마크의 구분은 Table 1.에 기술하였다. 추출된 각 랜드마크의 x, y 좌표는 이미지 너비와 높이를 기준으로 0과 1 사이로 정규화되었다.
- 데이터 분할
  학습 데이터는 Unreal Engine 데이터 12개 캐릭터 폴더 전체의 모든 이미지 + 각도 필터링된 BIWI 데이터의 90%, 검증 데이터는 각도 필터링된 BIWI 데이터의 10%의 데이터를 무작위로 샘플링 하여 검증 데이터를 생성하였다.
- YOLO 포맷 변환
  최종적으로 선택된 각 이미지에 대해 YOLO 포즈 추정 학습 형식에 맞는 라벨(.txt) 파일을 생성하였다. 각 라벨 파일은 [클래스_인덱스] [바운딩_박스_x_중심] [바운딩_박스_y_중심] [바운딩_박스_너비] [바운딩_박스_높이] [랜드마크1_x] [랜드마크1_y] ... [랜드마크23_x] [랜드마크23_y] 형식으로 구성되며, 모든 좌표는 0과 1 사이로 정규화되었다. 또한, 데이터셋의 경로와 클래스 정보, 랜드마크 개수(kpt_shape: [23, 2])를 명시하는 dataset.yaml 파일을 생성하였다.

Table 1. 23 Face Landmarks (0-22 Index)

Landmark Index	Spec
0	Nose Tip
1	Point slightly above nose tip
2	Center of nose bridge
3	Right eye inner corner
4	Right eye outer corner
5	Right eye top eyelid center
6	Right eye bottom eyelid center
7	Left eye inner corner
8	Left eye outer corner
9	Left eye top eyelid center
10	Left eye bottom eyelid center
11	Right mouth corner
12	Left mouth corner
13	Upper lip outer center
14	Lower lip outer center
15	Upper lip inner center
16	Lower lip inner center
17	Right eyebrow inner end
18	Right eyebrow center
19	Right eyebrow outer end
20	Left eyebrow inner end
21	Left eyebrow center
22	Left eyebrow outer end

Fig. 6. Locations of the 23 Facial Landmarks

2. Training Setup

안면 랜드마크 추출 모델과 안면 방향 추정 모델은 제안한 방법을 통해 구축한 학습 데이터셋으로 학습하였다. 안면 랜드마크 추출 모델의 경우 YOLOv11-pose 모델로부터 전이 학습을 진행하였으며, 안면 방향 추정 모델은 기존 모델[33]을 재학습하였다. 안면 랜드마크 추출 모델의 세부설정은 Table 2.와 같고 안면 방향 추정 모델은 Table 3.과 같다. 학습 환경은 아래 Table 4.와 같다.

Table 2. Training Hyperparameters of Facial Landmark Detection Model

Item	Setting
Model	yolov11n-pose
Input Image Size	640 x 640
Epochs	100
Batch Size	16
Optimizer	AdamW
Early Stopping	20 epochs

Table 3. Training Hyperparameters of Head Pose Estimation

Item	Setting
L2	0.00001
Dropout	15%
Epochs	100
Batch Size	32
Optimizer	Adam
Early Stopping	20 epochs

Table 4. System Environment

Category	Spec
CPU	AMD Ryzen 7 5800X 8-Core Processor
RAM	64GB
GPU	RTX3090 Turbo D6X 24GB
OS	Windows 10 Pro

3. Result of Facial Landmark Detection

학습은 patience=20으로 설정된 Early Stopping 조건에 따라 최대 100 Epoch가 아닌 43 Epoch에서 자동으로 중단되었다. 이는 약 23 Epoch 시점에서 검증 손실(Validation Loss)이 최적점에 도달한 후, 20 Epoch 동안 그 성능을 넘어서는 개선이 이루어지지 않았음을 의미한다. 이 조기 종료 기능에 따라 과적합이 방지된 최적의 모델이 최종 결과로 저장되었다. 주요 학습 결과는 Fig. 7. 검증 데이터 포즈 손실(Validation Pose Loss) 및 Fig. 8. 평균 정밀도(mAP50-95) 그래프를 통해 시각적으로 확인하였다.

최종 저장된 모델에 대해 원본 검증 데이터셋으로 정량적 평가를 수행하였다. 평가 결과, 안면 검출 성공률은 100%를 기록하여 모든 검증 데이터에 대해 누락 없는 안정적인 인식이 가능함을 확인하였다. 핵심점 추정의 정밀도를 나타내는 NME는 0.0083의 오차율을 보였으며, PCK@0.02 지표에서는 95.19%의 수치를 달성하였다. 이는 가상 환경의 3D 얼굴 모델링 데이터를 활용한 학습이 안면 핵심점 검출에 있어 우수한 성능의 정밀도를 확보하였음을 입증한다.

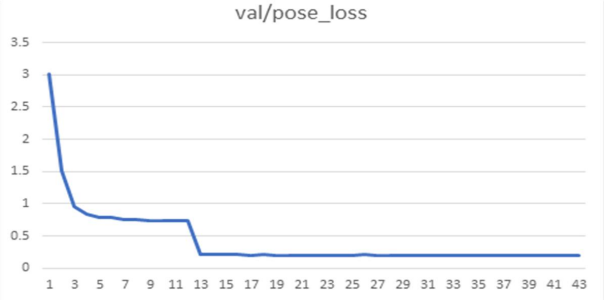

Fig. 7. Validation Pose Loss

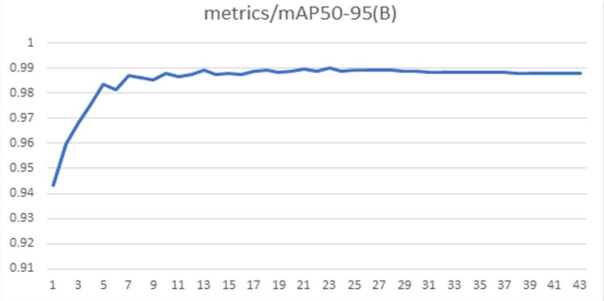

Fig. 8. Mean Average Precision (mAP50-95)

4. Data Ratio Comparison Experiment

본 절에서는 안면 방향 추정 모델의 최적 성능을 도출하기 위해 실제 데이터와 합성 데이터의 가장 효율적인 혼합 비율을 결정하는 실험을 수행하였다. 실제 데이터인 BIWI 데이터셋과 생성한 합성 데이터의 비율을 1(실제데이터 2,715개):1(합성데이터 3,375개)부터 1:12까지 조정하며 모델의 정밀도 변화를 측정하였다. 실험은 각 혼합 비율별로 동일한 하이퍼파라미터 환경에서 학습을 진행하였으며, MAE를 성능 평가 지표로 활용하였다. 실험 결과는 아래 Table 5.와 같다.

Table 5. MAE by Mixed Dataset Ratio

Mixing Ratio (Real:Synthetic)	Average MAE(°)
1:1	1.19
1:2	1.22
1:3	1.22
1:4	1.14
1:5	1.16
1:6	1.14
1:7	1.07
1:8	1.00
1:9	1.10
1:10	1.11
1:11	1.15
1:12	1.16

실험 결과, 합성 데이터의 비중이 증가함에 따라 전반적인 오차가 감소하는 경향을 확인하였다. 특히 실제 데이터 대비 합성 데이터의 비율이 1:8인 환경에서 1.00도의 가장 낮은 MAE를 기록하며 최적의 성능을 나타냈다. 이는 합성 데이터의 총량을 늘림에 따라 MAE가 1:1 비율에서 1.19도 1:8 비율에서 1.00도까지 점진적으로 개선됨을 확인하였고 특정 비율 이후 오차가 다시 증가하는 결과는 혼합 데이터 간의 정교한 균형이 모델 성능 향상에 기여함을 입증한다.

반면, 혼합 비율이 1:9를 초과하는 시점부터는 오차가 다시 소폭 상승하거나 정체되는 양상을 보였다. 이는 특정 임계치 이상의 합성 데이터 비중이 오히려 실제 데이터가 가진 고유한 특성을 희석시켜 Sim2Real Gap을 심화시킬 수 있음을 시사한다.

결론적으로 데이터 총량의 확대와 최적 비율의 설정은 실제 데이터 수집의 한계를 극복하고 Sim2Real Gap을 완화하는 요소임을 확인하였다. 따라서 본 연구에서는 가장 우수한 정밀도를 확보할 수 있는 1:8 비율을 최종 모델 학습을 위한 최적의 데이터 구성비로 채택하였다.

5. Result of Head Pose Estimation

5.1 Effectiveness of Hybrid Dataset Training

본 연구에서 제안하는 하이브리드 데이터셋의 유효성을 정량적으로 평가하기 위해 학습 데이터 구성에 따른 Ablation Study를 수행하였다. 모든 실험 조건은 동일하게 유지하였으며, BIWI 테스트 데이터셋을 대상으로 측정한 결과는 Table 6.과 같다.

Table 6. MAE by Training Dataset Configuration

Dataset	Value(MAE°)
Synthetic-only (Unreal)	13.23
Real-only (BIWI)	5.80
Proposed Mix	1.00

실험 결과, 가상 데이터만을 사용한 경우 13.23도의 높은 오차를 보였는데, 이는 가상과 실제 환경 사이의 도메인 차이인 Sim2Real Gap에 기인한다. 실제 데이터만을 사용한 경우 오차는 5.80도로 낮아졌으나, 학습 데이터 양의 부족으로 인해 신경망이 충분히 일반화되지 못하는 한계를 보였다. 반면, 제안하는 하이브리드 데이터셋은 대규모 합성 데이터의와 실측 데이터의 도메인 특징을 결합함으로써 1.00도의 가장 우수한 성능을 달성하였다. 이는 하이브리드 구성이 Sim2Real Gap을 완화하고 데이터 희소성 문제를 동시에 해결함을 증명한다.

5.2 Performance Comparison of Proposed and Existing Models

안면 방향 추정 모델 학습 결과로 각 축의 Roll, Pitch, Yaw 각도에 대한 Mean Absolute Error (MAE) 값은 Table 7.과 같다. 추가로 전체 평균 MAE를 기준으로 기존 연구들과 성능 비교를 진행하였다. 공정한 성능 평가를 위해 기존 모델들 모두 본 논문에서 제안하는 혼합 데이터셋을 사용해 새롭게 학습을 수행하였고 BIWI 테스트 데이터 셋의 near-frontal 영역을 대상으로 최종 성능 MAE 측정하였다. Table 8.은 실험의 결과를 보여준다. 실험 결과, 이미지 전체를 입력으로 사용하는 HopeNet, FSA-Net, TriNet, 6DRepNet은 각각 1.17도, 1.60도, 1.16도, 1.22도의 평균 오차를 기록하며 혼합 데이터셋을 통해 기존 연구들의 오차율도 향상되었으며 본 논문에서 제안 하는 모델이 1.00도로 기존 연구 모델들에 비해 성능이 우수함을 확인 할 수 있었다.

이러한 결과는 입력 데이터의 형태 차이에도 불구하고, 랜드마크 기반의 기하학적 특징 추출이 안면 방향 추정 작업에서 더 효과적임을 시사한다. 기존의 이미지 기반 모델들은 배경, 조명, 피부 질감 등 시각적 노이즈가 포함된 전체 이미지를 학습하므로 Sim2Real Gap의 영향을 직접적으로 받는다. 반면, 제안 모델은 환경 변화에 강건한 23개의 핵심 랜드마크 좌표만을 특징량으로 사용함으로써 실제 환경에서의 정밀도를 극대화하였다. 이는 동일한 혼합 데이터셋을 사용한 조건에서 특징 추출 방식의 차이가 모델의 최종적인 일반화 성능 향상에 결정적인 요인으로 작용하였음을 확인하였다.

Table 7. Pose Error Summary

Metric	Value(MAE°)
Roll	1.04
Pitch	1.18
Yaw	0.78
Overall MAE	1.00

Table 8. Test Results of Head Pose Estimation Models (BIWI)

Models	Average MAE(°)
HopeNet[9]	1.17
FSA-Net[10]	1.60
TriNet[11]	1.16
6DRepNet[12]	1.22
Our	1.00

6. Real-time Inference and Visualization

학습된 랜드마크 추출 모델(YOLO)과 안면 방향 추정 모델을 결합하여 실제 웹캠 환경에서의 성능을 확인하였다. YOLO 랜드마크 추론, Keras 각도 예측 및 시각화 과정을 포함한 전체 추론 파이프라인은 21.2 FPS의 평균 속도로 작동함을 확인하였다. 웹캠 영상에서 안면 방향 추정이 가시화된 영상에서 Fig. 9.는 정면, Fig. 10.은 왼쪽, Fig. 11.은 오른쪽을 바라보는 샘플 이미지를 보여준다.

Fig. 9. Real-time Pose Estimation (Frontal View)

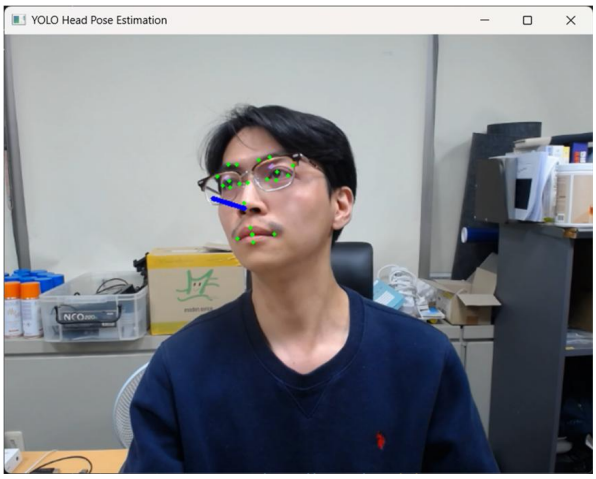

Fig. 10. Real-time Pose Estimation (Left-Facing View)

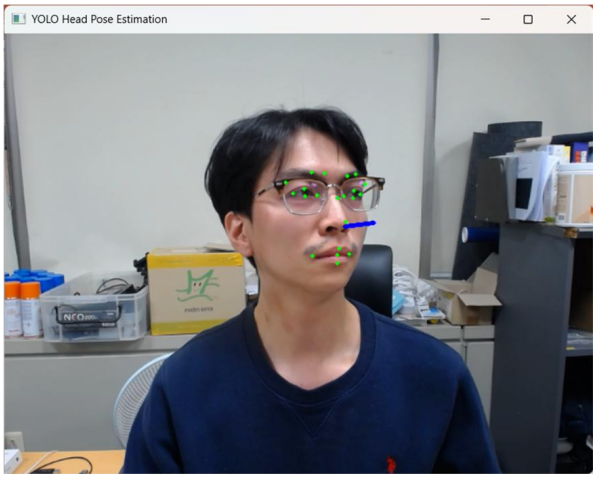

Fig. 11. Real-time Pose Estimation (Right-Facing View)

V. Conclusions

본 논문은 실제 데이터 수집의 한계를 극복하고, 안면 방향 추정 시스템의 성능을 향상시키기 위해 합성 데이터와 실제 데이터를 결합하여 학습 데이터베이스를 구축하고, 이를 기반으로 YOLOv11-pose 기반 얼굴 랜드마크 검출 모델 및 안면 방향 추정 모델의 학습 및 추론하는 방법을 제안하였다.

첫째, Unreal Engine(언리얼 엔진)과 MetaHuman(메타휴먼)을 활용하여 구축한 40,500장의 합성 데이터와 각도 필터링된 실제 BIWI 데이터 2,715장을 통합하여 총 43,215장의 학습 데이터베이스를 구축하였다. 특히 데이터 혼합 비율 실험을 통해 실제 데이터와 합성 데이터가 1:8의 비율일 때 최적의 성능이 도출됨을 확인하였으며, 이는 합성 데이터와 실제 데이터의 혼합과 데이터 총량의 증가가 실제 환경 추론의 정밀도를 향상시킴을 정량적으로 입증하였다.

두 번째 단계에서는 학습된 YOLOv11-pose 검출기를 이용하여 웹캠 영상에서 안면을 검출하고, 23개의 주요 얼굴 랜드마크 좌표를 추출하였다. 이 좌표들은 안면 방향 추정 모델의 입력으로 사용되기 위해 정규화를 거쳐 46차원의 입력 벡터로 변환되었다.

마지막 단계는 생성된 랜드마크 입력 벡터를 학습된 DNN 기반 안면 방향 추정 모델에 입력하여 최종적으로 Roll, Pitch, Yaw의 3축 각도를 실시간으로 예측하였다.

본 연구는 고품질의 합성 데이터와 실제 데이터를 전략적으로 결합하여 YOLO 기반 얼굴 랜드마크 검출기를 성공적으로 학습시켰으며, 이를 안면 방향 추정 모델과 결합

하여 3축의 전체 평균 MAE 1.00의 낮은 오차율과 평균 21.2 FPS 처리 속도를 보이며 ±21도의 정면 근접 영역에서 높은 정확도를 보여주었다. 향후에는 다양한 조명 환경 및 가려짐 상황에 대한 강건성을 확보하기 위해 데이터셋을 추가로 확장하는 연구 및 보다 넓은 영역의 각도를 대상으로 모델 고도화를 진행할 예정이다.

ACKNOWLEDGEMENT

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024(Project Name: Development of UX service technology based on new technology convergence content for enjoyment of cultural content by passengers in mobility, Project Number: RS-2024-00441262, Contribution Rate: 100%)

REFERENCES

[1] A. Ng, "MLOps: From Model-centric to Data-centric AI," DeepLearning.AI, https://www.deeplearning.ai/the-batch/a-chat-with-andrew-ng-about-data-centric-ai/

[2] A. Memon, A. A. Manjotho, Q. A. Arain, A. Sulaiman, N. Pirzada, M. S. Al Reshan, M. Alsulami, and A. Shaikh, "Lightweight CNN-based head pose estimation using heatmaps and anthropometric facial measures," ICT Express, Vol. 11, No. 5, pp. 914-918, Oct. 2025.

[3] G. B. Song, "Problems in the Use of AI Facial Recognition Technology and Criminal Investigation," Kyungchalbeop Yeongu (The Korean Journal of Police Law), Vol. 22, No. 1, pp. 209-235, Feb. 2024.

[4] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, Vol. 6, No. 1, pp. 60, July 2019. DOI: 10.1186/s40537-019-0197-0

[5] Y. Zhu, C. F. Dagan, D. Ramanan, and A. L. Yuille, "Face Alignment Across Large Poses: A 3D Solution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 146-155, Las Vegas, USA, June 2016. DOI: 10.1109/CVPR.2016.23

[6] I. Nikolenko, "Synthetic Data for Deep Learning," Apress, pp. 1-200, 2021.

[7] Epic Games, Inc., "Unreal Engine," https://www.unrealengine.com

[8] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for Data: Ground Truth from Computer Games," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 102-118, Amsterdam, The Netherlands, Oct. 2016.

[9] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-Grained Head Pose Estimation Without Keypoints," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2074-2083, Salt Lake City, USA, June 2018. DOI: 10.1109/CVPRW.2018.00265

[10] Y. Yang, Z. Liu, X. Liu, and X. Wang, "FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation from a Single Image," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1087-1096, Long Beach, USA, June 2019.

[11] Z. Cao, Z. Chu, D. Liu, and Y. Chen, "A vector-based representation to enhance head pose estimation," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1188-1197, Waikoloa, USA, Jan. 2021. DOI: 10.1109/WACV48630.2021.00124

[12] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6DRepNet: A 6D Rotation Representation for Unconstrained Head Pose Estimation," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3671-3680, Montreal, Canada, Oct. 2021. DOI: 10.1109/ICCV48922.2021.00366

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, https://arxiv.org/abs/1704.04861

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6848-6856, Salt Lake City, USA, June 2018.

[15] N. Dhingra, "LwPosr: Lightweight Efficient Fine Grained Head Pose Estimation," arXiv:2205.09113, https://arxiv.org/abs/2205.09113

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, Las Vegas, USA, June 2016.

[17] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why Reinventing a Face Detector," arXiv:2105.12931, https://arxiv.org/abs/2105.12931

[18] C. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), pp. 7464-7475, Vancouver, Canada, June 2023.

[19] D. Y. Jung, H. J. Jung, and N. H. Kim, "Analysis of Fire Detection Performance of Deep Learning-based Yolo v7 and Yolo v8 Models using GAN Algorithm," Journal of Information Technology and Architecture, Vol. 14, No. 1, pp. 13-21, Mar. 2024. DOI: 10.22733/JITAE.2024.14.01.002

[20] Ultralytics, "YOLO11: Real-Time Object Detection and Pose Estimation," https://docs.ultralytics.com

[21] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real Time Head Pose Estimation from Consumer Depth Cameras," in Pattern Recognition (DAGM 2011), Lecture Notes in Computer Science, Vol. 6835, pp. 100-109, Berlin, Germany, 2011. DOI: 10.1007/978-3-642-23123-0_11

[22] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge," in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397-403, Sydney, Australia, Dec. 2013.

[23] E. Wood, T. Baltrusaitis, L. P. Morency, P. Robinson, and A. Bulling, "Learning an Appearance-based Gaze Estimator from One Million Synthesised Images," in Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA), pp. 131-138, Charleston, USA, March 2016.

[24] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces," in Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH), pp. 187-194, Los Angeles, USA, Aug. 1999.

[25] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401-4410, Long Beach, USA, June 2019.

[26] G. L. T. Cunha, F. G. De Souza, and D. B. D. S. Junior, "HeadNeRF: A Real-time NeRF-based Head Model," arXiv:2112.05637, https://arxiv.org/abs/2112.05637

[27] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," Journal of Machine Learning Research, Vol. 17, No. 59, pp. 1-35, 2016.

[28] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2223-2232, Venice, Italy, Oct. 2017.

[29] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial Discriminative Domain Adaptation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7167-7176, Honolulu, USA, July 2017.

[30] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from Simulated and Unsupervised Images through Adversarial Training," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2242-2251, Honolulu, USA, July 2017.

[31] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 23-30, Vancouver, Canada, Sep. 2017.

[32] J. Choi, J. Kim, I. Choi, and K. Seo, "The photo-realistic facial landmark dataset created with scan-based game characters," Journal of Digital Contents Society (J. DCS), Vol. 23, No. 11, pp. 2259-2268, Nov. 2022.

[33] U. Kim, H. Kim, S. Chun, J. Yun, and S. Hong, "AI-Based Facial Orientation Estimation Framework in a Virtual Occupant Environment," 2025 Summer Conference of Society for Computational Design and Engineering (CDE), pp. 1-2, Pyeongchang, Korea, July 2025.

Authors

Un-Yong Kim received the B.S. degree in Computer Science from Honam University in 2018. He is currently a research scientist at the Korea Photonics Technology Institute (KOPTI).

Interests: Image Processing, Machine Learning.

Sungkuk Chun received his B.S. and M.S. degrees in Media Engineering from Soongsil University, Seoul, Korea, in 2009 and 2011, respectively, and earned his Ph.D. in HCI and Robotics from the University of Science and Technology (UST), Korea, in 2017. He is currently a senior research scientist at the Korea Photonics Technology Institute (KOPTI). His research interests include computer vision, artificial intelligence-based human motion analysis, and natural user interaction.

Jeongrok Yun received his M.S. degree of Science in Electrical Engineering and Computer Science from Chonnam National University in 2019. He is currently a research scientist at the Korea Photonics Technology Institute (KOPTI). Interests: Image Processing, Machine Learning.

Ju-Yeong Park received the M.S. degree in Data Science from Chonnam National University in 2025. He is currently a research scientist at the Korea Photonics Technology Institute (KOPTI).

Interests: Image Processing, Machine Learning.

Sung-Hoon Hong received Ph.D. degree in Electronic Engineering from KAIST in 1999. He is currently a professor at Chonnam National University. His research interests include computer vision, AI based image processing and semiconductor design.

Hoe-Min Kim received a B.S. in mechanical engineering from Chung-Ang University, Korea in 2002, and M.S. and Ph.D. degree in mechatronics at GIST in 2004 and 2012, respectively.

He is currently a principal researcher of Korea Photonics Technology Institute (KOPTI) in Gwangju, Republic of Korea. His research interests include realistic image synthesis, AI and AR/VR.