한국음향학회지 제43권 제5호 pp. 529~535 (2024)
The Journal of the Acoustical Society of Korea Vol.43, No.5 (2024)
https://doi.org/10.7776/ASK.2024.43.5.529

pISSN : 1225-4428
eISSN : 2287-3775

이머시브 오디오 패닝을 위한 깊이 정보 기반 객체 추적 및 무대 크기 예측에 관한 연구

Research on depth information based object-tracking and stage size estimation for immersive audio panning

이강은,[1] 박홍준,[1] 김성영[1,2†]

(Kangeun Lee,[1] Hongjun Park,[1] and Sungyoung Kim[1,2†])

[1]한국과학기술원, [2]로체스터 공과대학교

(Received July 2, 2024; revised September 9, 2024; accepted September 12, 2024)

초 록: 본 논문은 미디어 컨텐츠 제작을 위한 자동 오디오 패닝 기술 구현에 관한 연구 내용을 다룬다. 이전까지, 오디오 오브젝트를 지속적으로 추적하는 것은 사람의 수동 작업에 의존하였다. 이머시브(몰입형) 오디오의 시대가 도래함으로써, 자동 오디오 패닝 시스템의 필요성은 점차 부각되었지만, 현재까지 현업에 적용한 연구까지는 진행되지 않고 있다. 이에 본 연구팀은 시청각 조화를 고려한 깊이 정보 기반 객체 추적을 적용한 자동 오디오 패닝 시스템을 제안한다. 시스템은 먼저 2차원의 좌표를 기반으로 깊이 정보를 계산하여 이를 반영한 3차원의 Top-View 시점 변환을 모델링한다. 또한, 실제 무대 공간의 이미지를 입력 값으로 받아, 무대 바닥의 가로 및 깊이를 예측하는 모델을 적용한다. 무대 크기를 예측한 값이 시점 변환에 적용되기에 별도의 깊이 데이터 학습이 추가적으로 요구되지 않는다. 본 연구에서 제안하는 시스템 유효성을 검증하기 위해 Unity 기반의 샘플 비디오를 사용하여 파일럿 테스트를 진행했다. 본 시스템은 많은 오디오 엔지니어들에게 자동화된 오디오 패닝 기능을 제공함으로써 오디오 프로덕션의 작업 효율 개선에 도움을 줄 것으로 예상된다.

핵심용어: 오디오 패닝, 객체 추적, 깊이 정보, 무대 크기 예측, 컴퓨터 비전

ABSTRACT: This paper presents our research on automatic audio panning for media content production. Previously, tracking an audio was done manually. With the advent of the immersive audio era, the need for an automatic audio panning system has increased, yet no substantial research has been progressed to date. Therefore, we propose a computer vision-based human tracking and depth feature processing system which processes depth feature through using 2-dimensional coordinates and models 3-dimensional view transformation for automatic audio panning to ensure audiovisual congruence. Also, this system applies stage size estimation model which gets input as an image and extrapolates stage width and depth as meter unit. Since our system estimates stage sizes and directly applies them to view transformation, no additional depth data training is required. To validate the proposed system, we also conducted a pilot test with Unity based sample video. Our team expects that our system will enable automated audio panning, assisting many audio engineers.

Keywords: Audio panning, Object tracking, Depth information, Stage size estimation, Computer vision

PACS numbers: 43.55Ka, 43.58Ta, 43.60Dh

† **Corresponding author:** Sungyoung Kim (sungyoung.kim@kaist.ac.kr)
Korea Advanced Institute of Science and Technology N25 2F #3221, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
(Tel: 82-42-350-2917, Fax: 82-42-350-2910)

"이 논문은 2024년도 한국음향학회 춘계학술대회에서 발표하였던 논문임."

I. 서 론

음향은 물체의 울림이 만들어내는 소리의 생성, 이동, 환경과의 상호작용 및 최종 전달의 전체적인 총괄하는 복합적인 현상이다. 기술의 발전에 따라 소리를 기계적으로 생성하고 전달하는 영역을 오디오로 구분하게 되었고, 기술의 발전에 따라서 이 오디오 기술 역시 빠르게 진일보 하고 있다. 다양한 오디오 기술 중에서 몇 가지의 오디오 채널을 사용하느냐는 청취자들이 얼마나 더 실감나는 오디오 경험을 할 수 있는가를 결정한다. 이러한 채널 기반의 오디오는 큰 범주 내에서 다음의 다섯 종류로 구분될 수 있다.

1. 모노
   - 하나의 스피커를 이용해서 소리를 청취하며 단일 채널 오디오를 의미한다.
2. 스테레오 오디오
   - 좌우 구분을 기준으로 총 두 개의 개별 채널로 청취자의 전면에서 음원의 위치를 자유롭게 제어한다.
3. 서라운드 사운드
   - 다섯 개의 채널을 통해 청취자를 감싸는 수평면에자유롭게 음원의 위치를 제어한다.
4. 3D 오디오
   - 서라운드 사운드를 수직적으로 확장하여 청취자가 위치한 공간 어디에나 음원의 위치가 가능하도록 하는 이론적으로 완성된 오디오 시스템을 의미한다.
5. 바이노럴 오디오
   - 더미헤드 마이크를 사용해 3D 오디오 신호를 녹음하고 헤드폰을 통해서 청취하는 시스템을 나타낸다. 양 귀에 신호를 전달하기에 스테레오 오디오와 채널 수는 동일하다.

이러한 다양한 오디오 채널안에서 재생되는 음원의 위치를 정하는 신호 처리가 패닝이다. 전통적으로 패닝은 모노 오디오를 스테레오 공간에 위치하는 것을 지칭하며 두 채널 신호의 레벨 혹은 시간 차이를 조절하는 방법을 통해 음원의 위치를 제어한다.[1]

Fig. 1. (Color available online) Example image of stereophonic panning function inside of a Digital Audio Workstation (DAW) program. The left image shows the panning hard left and the rigth image shows the opposite with corresponding gain strength (light blue color).

Fig. 1은 스테레오에서 좌우로 패닝하는 경우의 예를 보여준다. 이렇게 좌-우 방향성을 부여하는 패닝 작업은 현재까지 오디오 엔지니어의 수작업의 영역으로 남아있다. 음원의 수가 적고 스테레오 오디오와 같이 제어할 공간이 정면상의 두 스피커를 연결하는 선 안으로 한정되는 경우 이러한 수작업이 큰 부담이 아닐 수 있으나, 서라운드(면) 그리고 3D 오디오(공간)에 있어서의 패닝은 작업의 난이도를 지수적으로 증가시킨다. 다양한 장소에 다양한 음원을 배치하는 일은 숙련된 엔지니어에게도 무척 시간과 노동집약적인 업무가 되고 있다. 이를 다른 흐름에서 보면, 영화 산업, 게임 엔터테인먼트, 콘서트 등의 분야처럼 오디오와 시각적 정보들 간의 긴밀한 상관관계 해석을 필요로 하는 분야에서 있어서는 개별 엔지니어의 능력에 따라서 주어진 시청각 정보의 일치가 일관성있게 재현되지 못할 수 있다는 문제를 가져오기도 한다.

이에 새로운 패닝 기법, 즉 개별 엔지니어의 시청각 정보 인지를 바이패스하여 주어진 음원의 위치 정보를 오디오 시스템이 자동으로 반영하는 패닝 기법의 필요가 부각되고 있다. 이러한 기법은 기술적으로 비디오 정보에서 음원의 위치 정보를 인식하고 동시에 그 음원의 움직임을 추출해 오디오에 매핑하는 방식을 택한다.

본 연구팀은 앞서 언급된 내용을 기반으로 시청각 조화를 이룰 수 있는 새로운 접근 방법 기반의 패닝 시스템을 제안한다. Fig. 2는 해당 시스템의 전체적

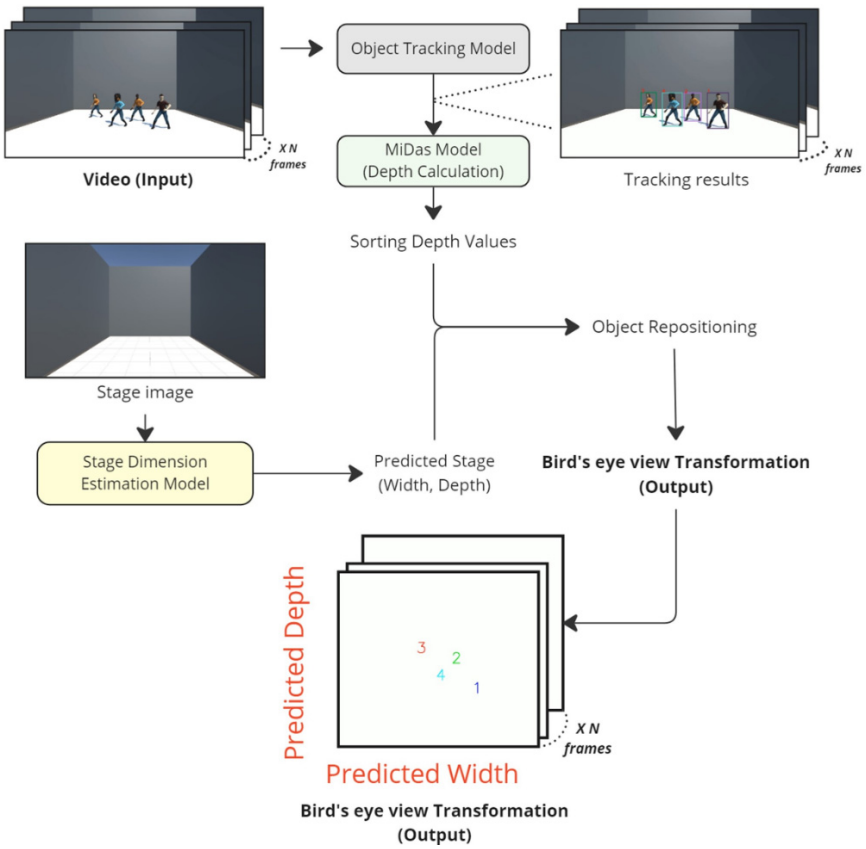

Fig. 2. (Color available online) Overview of the proposed system.

인 동작 구조와 데이터 흐름을 시각화한다. 해당 시스템은 세 가지의 연구적 기여 요소를 지닌다. 첫째, 이전 논문들에서는 시도되지 않았던 깊이 정보를 객체 추적 결과와 결합하여 시점 변환을 처리하는 자체 알고리즘을 제안한다. 둘째, 더욱 몰입감을 증대하기 위해 입력 영상의 무대 크기를 예측하는 모델을 제안한다. 무대 크기 정보는 가로와 세로(깊이) 길이이며 파인튜닝을 적용하여 최소한의 학습 과정을 통해 최대의 성능 도달을 목표했다. 셋째, 본 시스템에서 객체 재배치 알고리즘을 도입하여 앞서 처리된 데이터를 기반으로 예측된 무대 크기를 기준으로 객체 위치를 다시 계산하여 시각화 한다.

본 논문은 다음과 같이 구성된다. II에서는 객체 추적 적용을 위한 내용을 설명할 것이고, III에서는 깊이 정보 추출 및 활용에 초점을 맞춘다. IV에서는 무대 크기 예측 및 자세한 방법론에 대해 기술할 예정이며, 마지막으로 결론 및 추후 연구 방향성에 대해 설명한다.

II. 다중 객체 추적

컴퓨터 비전은 시각적 입력 값에서 유의미한 특징 값을 추출하여 신경망에서 학습시켜 처리하는 모든 과정을 지칭한다. 세부적으로는 객체 인식 / 분류 및 추적, 이미지 분류 및 세그멘테이션, 얼굴 인식, 추적 등의 대표적 예시들이 존재한다.

앞선 내용과 같이, 본 연구팀은 시청각 조화에 기반한 자동 오디오 패닝 시스템을 제안하기에, 여러 예시 중에서도 객체 추적에 초점을 맞추어 연구 방향을 진행하였다. 이를 위해 시각적 데이터 즉, 비디오 데이터가 입력 값으로 주어졌을 경우, 이때 다중 객체 추적(Multi Object Ttracking, MOT)을 적용하였다. 현재까지 이전 연구에서 제시된 다양한 객체 추적기가 있기에, 본 연구팀은 기존 연구들 중에서 5편의 논문 후보군을 생성하고 사전 연구 조사 및 전체 코드 리뷰 작업을 진행했다. 또한 유효성 검증을 위해, 사전 학습 된 공식 체크포인트를 Github에서 가

져와 전반적인 일반화 성능을 확인했다.

Zhang *et al.*[2]이 제안한 ByteTrack 모델은 각 특징 간의 상관관계를 파악하는 BYTE 라는 자체 알고리즘을 기반으로 높은 성능의 다중 객체 추적을 수행한다. 해당 논문의 코드를 리뷰하고 랜덤 영상을 기반으로 모델 성능을 평가할 수 있었다.

Sun *et al.*[3]에서는 DanceTrack 이라는 이름으로 다수의 사람들이 동일한 의상과 비슷한 행동을 반복적으로 수행하는 대규모 영상 데이터 셋을 구축하였다. 또한, Reference [4]를 기반으로 depth를 계산하는 추가 모델도 학습하여 적용하였으며 이를 통한 top view 에서 영상을 변환한 결과를 얻어낼 수 있었다.

Zhang *et al.*[5]은 Motrv2라는 모델을 제안했다. 해당 논문은 Ge *et al.*[6]의 결과를 추가적인 값으로 활용하여 전반적인 모델 성능을 높였으며, 두 개의 Proposal 쿼리, Tracking 쿼리 기반으로 더욱 정확도를 개선하였다.

Cao *et al.*[7]은 OC-SORT 라는 모델을 제안했다. 우리는 저자들이 제공하는 모델 체크포인트를 기반으로 Unity 환경의 더미 객체들이 동일한 동작으로 춤을 추고 있는 영상의 객체 인식 및 추적 결과는 Fig. 3에서 확인할 수 있었다.

하지만, 객체 추적에서 극복하기 어려운 문제점들 중 하나가 “ID Switching”이다. 이는 영상에서 객체 간의 가려짐 현상 이후 다시 추적을 하는 경우, 서로 간의 고유 ID 번호가 바뀌어 버리거나 새로운 ID로 매핑이 되어버리는 문제를 지칭한다. 이러한 문제는 객체 추적에서 치명적인 성능 저하를 유발한다. 본 연구에서 제안하고자 하는 시스템은 객체 추적 결과를 기반으로 이후 단계를 진행하기에 해당 단계가 중요한 첫 걸음이라고 말할 수 있다. 그러므로 실험 중 가장 “ID Switching” 문제가 가장 적게 발생한 OC-SORT모델[7]을 객체 추적기로 선택했고 전체적인 코드 리뷰를 진행하여 추후 연구에 필요한 작업을 준비했다.

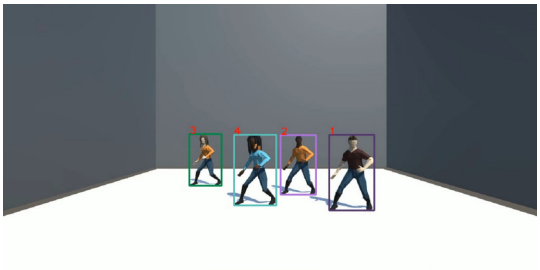

Fig. 3. (Color available online) Visualization of OC-SORT in Unity.

Ⅲ. 깊이 정보 추출

이전 단계에서의 객체 추적이 완료되고, 이제는 시각 데이터의 깊이 정보를 추출하기 위해서 별도의 깊이 정보 추출 모델을 적용해야 한다. 본 논문에서 다루는 시각 데이터는 여러 프레임의 집합인 영상 데이터이다.

각 영상프레임 별, 깊이 특징 맵을 추출하기 위해서는 다양한 단안의 깊이 특징 추출기(Monocular Depth feature extractor) 모델을 고려해야 한다. Ranftl *et al.*[8]을 선택했고 이는 MiDas 라고 불린다. 해당 모델은 Pytorch 기반의 패키지 형식으로 코드에서 호출하여 사용할 수 있다. 편리함과 코드 이식성이 좋기에 사용하였고, 세 가지의 모드 중 hybid 모드를 선택했다.

본 연구에서는 구현 단계에서 MiDas[8] 모델을 기반으로 각 영상 프레임 별 깊이 데이터를 추출해서 시각화하는 코드를 구현했고 실험을 진행했다. Fig. 3에 사용된 프레임과 동일한 것을 기반으로 깊이 정보를 시각화한 결과가 Fig. 4에 나타나있으며, 전방에 가깝게 위치할 수록 더욱 밝은 빛으로 시각화 된다는 것을 확인할 수 있다.

다음 단계는 각 프레임별로 추출된 두 개의 결과물(객체 추적 시각화, 깊이 정보 시각화) 을 결합하는 것이다. 다시 말해, 객체 추적 시각화 결과 값에서 나온 객체 ID 값, 객체 인식한 bounding box 좌표 값들

Fig. 4. (Color available online) Depth feature map visualization of Fig. 3.

을 기반으로 깊이 정보를 잘라낸다. 최종적으로 사람이 인식된 결과를 기반으로 그에 맞게 잘려진 깊이 정보가 결과물로 나온다.

IV. 무대 크기 예측 및 객체 재배치

자동 오디오 패닝 시스템 구축을 위해서 실제 무대 크기 예측이 선행되어야 한다. 본 연구에서는 무대 크기 예측을 위해 파인튜닝 기법 적용과 이미지 입력 값의 변화를 통한 성능 개선에 초점을 맞추어 연구를 진행했다.

본 연구팀은 파인튜닝 기법을 기존에 학습이 된 He *et al.*[9] 모델 구조에 적용했다. 해당 모델은 이미지 분류에 맞게 제시된 모델이며, ResNet이라고 불린다. 본 연구에서는 해당 모델이 18개의 레이어로 구성된 버전을 사용했다.

ResNet[9] 신경망 모델은 Fig. 5의 내용과 같이 잔차 연결을 적용하여 다음 레이어의 학습에 반영했다. 지속적인 잔차 연결을 통해 신경망 학습에서 발생할 수 있는 기울기 소실 문제 해결을 위해 도입했고 더욱 안정적인 모델 학습을 가능하게 했다.

위와 같이 사전 학습된 ResNet 모델을 로드하여 파인튜닝을 적용한다. 파인튜닝을 본 연구에 도입한 이유는 새로운 도메인에 대한 성능을 향상할 수 있는 방법이며, 소량의 데이터를 기반으로도 좋은 성능을 얻을 수 있는 방법이기에 도입하였다. 다시 말해, 기존의 이미지 분류 작업에서 본 연구의 목적에 맞는 무대 크기 예측 작업으로의 도메인 변경이 있기에, 새로운 무대 이미지 데이터 구축과 더불어 파인튜닝을 적용하였다.

파인튜닝을 진행하기에 앞서, 본 연구팀은 무대 이미지 및 무대 크기 정보가 매핑된 데이터 셋을 새롭게 구축했다. 해당 데이터 셋은 기존에 존재하지 않은 데이터 셋이다. 그래서 무대 이미지와 그에 상응하는 크기 정보들을 수작업으로 모았고 검수를 통한 소량의 데이터 셋을 구축하였다. Fig. 6은 자체 수집한 데이터 셋의 예시이며 하나의 무대 이미지와 상응하는 무대 가로 길이, 세로(깊이) 길이를 파일 이름으로 함께 저장하여 추후 개발 과정에서 데이터 전처리에 용이하도록 했다.

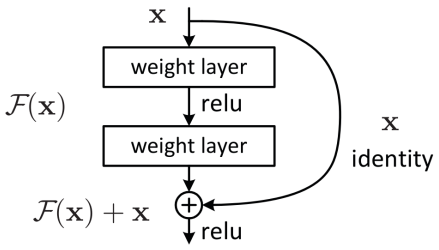

Fig. 5. Residual connection based learning.

데이터 구축이 완료되었기에, 파인튜닝 작업을 진행하며 전체적인 과정은 Fig. 7로 확인할 수 있다. 적은 데이터 셋으로 파인튜닝을 진행하며 학습을 위한 세팅으로 Epoch은 20으로 설정했다.

파인튜닝을 통한 최대 성능 도달을 위해, 본 연구팀은 총 세 종류의 입력 값 변화를 적용했다. 첫번째, 원래의 무대 이미지 자체를 입력 값으로 주는 것이다. 두번째, 본래의 무대 이미지에서 추출된 외곽선 특징 값을 입력 값으로 주는 것이다. Canny[10] 알고리즘을 기반으로 추출된 특징 값을 입력 값으로 주는 것이 마지막 세번째에 해당한다.

또한, 파인튜닝에서 가장 빈번하게 발생하는 문제가 바로 과적합 문제이다. 본래 파인튜닝이 소량의 데이터 셋을 기반으로 진행되기 때문에 과적합이 잘 발생할 수 있고, 이론 인해 모델 성능 또한 현저하게 감소할 가능성이 존재한다.

그래서 본 연구에서는 이러한 위험성을 최대한 줄이기 위해, 데이터 증강을 적용했다. 현재 연구팀이 구축한 무대 데이터 셋은 하나의 무대 이미지와 그에 맞는 무대 가로 길이 그리고 세로(깊이) 길이로 구성된다. 무대의 크기 정보는 가변적이지 않기에, 다양한 각도에서 동일한 무대를 촬영한 이미지들을 추가하여 다각도에서의 무대 이미지 특징 값들을

Fig. 6. (Color available online) Example images of custom dataset.

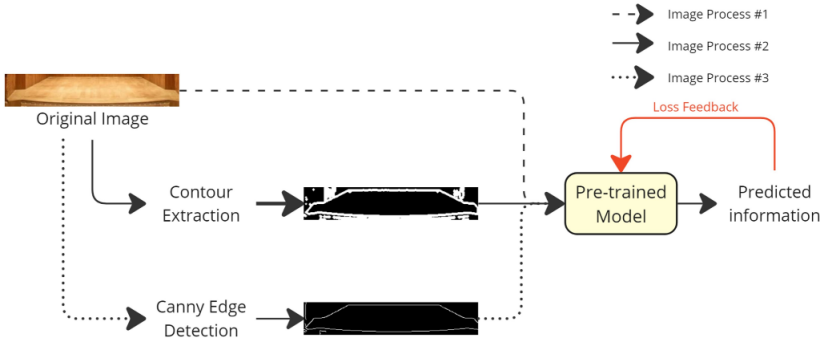

Fig. 7. (Color available online) Overview of fine-tuning process.

Algorithm 1 Object Repositioning

Input: Video frames F
Output: Bird-view like video transformation

$L \leftarrow$ (Max(depth), tracking ID)
$image \leftarrow$ Front-view of stage
$width, depth \leftarrow SSE(image)$ ▷ Stage Size Estimation model(fine-tuned)

for $frame$ in F **do**
  $depthMap \leftarrow MiDas(frame)$ ▷ Depth Calculation Model
  $bboxes \leftarrow OC\_SORT(frame)$ ▷ Returns n bounding boxes with IDs
  for box in $bboxes$ **do**
    $cropped\_depth \leftarrow$ CROP$(depthMap, box)$
    $D_{max} \leftarrow$ MAX$(cropped_depth)$
    $new_y \leftarrow$ MAP(D_{max}) ▷ new_y : coordinate for bird's eye-view
    $L.append(D_{max}, box\_ID, new_y)$
  end for
  SORT$(L,$ key $\leftarrow D_{max})$ ▷ Sort L based on D_{max}
  SHOW$(frame, L, width, depth)$ ▷ Visualize results
end for

Fig. 8. Pseudo code of the proposed algorithm.

추출하여 파인튜닝 과정에 적용하려 했고 파인튜닝을 진행했다.

다음 단계로는 객체 재배치를 진행한다. 객체 재배치는 front-view를 top-view로의 시점 변환을 가능케 하는 알고리즘을 제안한다. 해당 알고리즘의 의사 코드는 Fig. 8에서 확인할 수 있다.

해당 알고리즘은 영상을 입력 값으로 받은 경우, 프레임 하나씩 처리하면서 결과를 시각화한다. 우선, 개별 프레임 별로 MiDas[8]를 깊이 정보 추적기로 이용하여 깊이 정보를 추출한다. 다음으로 각 프레임 별로 OC-SORT[7] 객체 추적기를 이용하여 개별 ID와 bounding box 좌표를 추출한다.

이때, 여러 개의 bounding box 좌표들이 나올 수 있기에, 반복문을 순회하면서 각 bounding box 좌표 정보에 맞게끔 깊이 정보를 잘라낸다. 이후, 잘라낸 깊이 정보를 각 객체의 ID에 맞게 저장한다. 이렇게 저장된 정보는 정렬되고, 해당 데이터를 기반으로 각 개체간의 상대적 깊이 순서를 파악할 수 있게 된다.

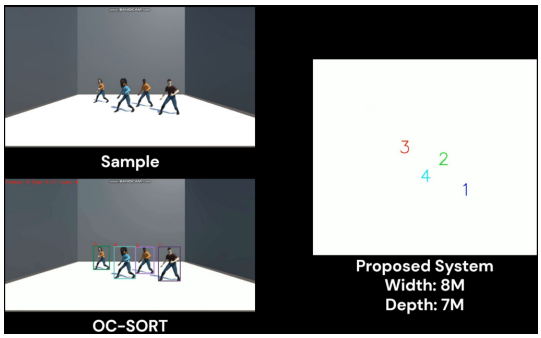

Fig. 9. (Color available online) Visualization of the final output.

파악된 객체 간의 상대적 깊이 특징 값들을 기반으로 무대 크기 예측 모델이 예상한 무대의 가로 길이, 세로(깊이) 길이에 맞는 평면에 각 객체들의 좌표들을 재배치한다. 재배치 과정에서 x 좌표는 무대 가로 길이에 맞게끔 변환이 되고 y 좌표는 앞서 추출된 깊이 정보에 기반하여 무대의 세로(깊이) 길이에 알맞게 변환되어 시각화 된다.

더욱 안정화된 최종 결과를 위해, 좌표 데이터를 다듬고 정리하는 작업을 추가했다. 여러 방법들 중 하나인 이동 평균 법을 적용하여 전체적인 좌표 데이터를 더욱 부드럽게 표현될 수 있도록 정리하는 작업을 추가했다. 그 결과 이전 보다 더욱 자연스러운 시점 변화 결과를 얻을 수 있었으며, 최종 영상 결과는 Fig. 9에서 확인할 수 있다.

V. 결 론

본 연구에서는 기존의 오디오 엔지니어들의 업무적 복잡도를 완화하는 것에 도움을 줄 수 있는 자동

화된 오디오 패닝 시스템을 새롭게 제안한다. 객체의 위치 및 움직임을 컴퓨터 비전의 다중 객체 추적 기술을 사용하여 파악한다. 그리고 나서 각 영상 프레임 마다 추출될 수 있는 깊이 정보를 얻게 되며, 새롭게 구축한 무대 크기 예측 모델과 재배치 알고리즘을 통해 최종적으로는 정면 시점에서 상면 시점으로 변환하는 결과를 얻게 된다.

결론적으로 본 연구는 총 세 가지 기술적 측면에 기여한다. 첫째, 시청각 조화를 목표로 한 자동화된 오디오 패닝 시스템을 처음으로 제안했다. 둘째, 파인튜닝을 통한 무대 크기 예측 모델 확보 및 그에 필요한 무대 데이터를 직접 수집 및 전처리 과정을 진행했다. 셋째, 깊이 정보와 무대 크기 예측 데이터만으로도 시점 변화의 결과를 얻을 수 있었다.

이러한 기술적 기여를 기반으로, 앞으로 본 연구팀은 더욱 발전된 알고리즘을 개발 및 제안할 예정이며, 어떤 상황에서나 적용될 수 있는 강건하며 시청각 조화를 목표로 한 자동화된 오디오 패닝 시스템을 제안하는 것을 추후 계획으로 설정한다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00222383).

References

1. J. Klapholz, “Fantasia: Innovations in sound,” J. Audio Eng. Soc. **39**(1/2), 66-70 (1991).
2. Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, “Bytetrack: Multi-object tracking by associating every detection box,” Proc. ECCV, 1-21 (2022).
3. P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, and P. Luo, “Dancetrack: Multi-object tracking in uniform appearance and diverse motion,” Proc. IEEE Conf. CVPR. 20993-21002 (2022).
4. A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The kitti vision benchmark suite,” Proc. IEEE Conf. CVPR. 3354-3361 (2012).
5. Y. Zhang, T. Wang, and X. Zhang, “Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors,” Proc. IEEE Conf. CVPR. 22056-22065 (2023).
6. Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “Yolox: Exceeding yolo series in 2021,” arXiv preprint arXiv:2107.08430 (2021).
7. J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, “Observation-centric sort: Rethinking sort for robust multi-object tracking,” Proc. IEEE Conf. CVPR. 9686-9696 (2023).
8. R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” IEEE Trans. Pattern Anal. Mach. Intell. **44**, 1623-1637 (2020).
9. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE Conf. CVPR. 770-778 (2016).
10. J. Canny, “A computational approach to edge detection,” IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-8**, 679-698 (1986).

저자 약력

▸ 이 강 은 (Kangeun Lee)

2021년 2월 : 서경대학교 컴퓨터과학과 학사
2021년 3월 ~ 2023년 8월 : 고려대학교 인공지능대학원 인공지능학과 석사 수료
2023년 8월 ~ 2024년 1월 : 한국과학기술원 문화기술대학원 AIRIS LAB 인턴
2024년 2월 ~ 현재 : 한국과학기술원 문화기술대학원 석사

▸ 박 홍 준 (Hongjun Park)

2018년 2월 ~ 현재 : 한국과학기술원 전기및전자공학부 학사

▸ 김 성 영 (Sungyoung Kim)

1996년 2월 : 서강대학교 전자계산 학사
2004년 6월 : McGill대학교 Sound Recording 석사
2009년 2월 : McGill대학교 Sound Recording 박사
1995년 12월 ~ 2001년 6월 : 한국방송공사 라디오 기술국
2007년 7월 ~ 2012년 12월 : Yamaha Corp. R&D 연구원
2012년 8월 ~ 현재 : RIT 부교수
2022년 12월 ~ 현재 : 한국과학기술원 부교수