Enabling X-ray vision via multi-camera fusion for AR applications

I. Pastaltzidis, I. Karakostas, N. Dimitriou, D. Tzovaras

Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece.

Corresponding authors: gpastal@iti.gr, iason@iti.gr, nikdim@iti.gr, dimitrios.tzovaras@iti.gr

Keywords: Augmented Reality, X-Ray vision, 3D pose estimation

Introduction

While Augmented Reality (AR) enriches our Field-of-View (FoV) with virtual objects and annotations, it can also be used to enhance the users' situational awareness and perception (Apostolakis et. al, 2022). A useful application of AR is to make users aware of completely occluded objects. This work aims to address this type of problem, i.e. detecting occluded occupants of a building, using RGB(D) cameras and human 3D pose estimation towards annotating them on AR glasses. Estimating the 3D pose from a multi-view camera setup is a topic that has attracted a lot of attention in recent years. Multi-view, multi person 3D pose estimation in general lacked in speed when compared to single-view methods. The method introduced by Ye et. al (2022) has made great strides towards real-time multi-view pose estimation and is exploited in the proposed pipeline. The projection of totally occluded 3D poses on AR glasses, enables a form of X-ray vision, since the AR user can see fully occluded occupants inside the building even under motion. A short video demonstration of the pipeline is provided in this link. This application targets first responders (police, paramedics, security guards) as its users and can be deployed in facilities that have CCTV system (malls, airports, etc.). The proposed system is designed for scenarios where obstacles obstruct the line of sight, making it challenging for users to observe what is behind them. It proves particularly valuable in situations where there's a critical need for real-time awareness of the posture and location of concealed individuals. For instance, in law enforcement during a security incident, police officers require precise information about the posture and position of room occupants before taking action. Similarly, such spatial information would be very useful for first responders, like firefighters, when entering spaces during emergencies.

X-ray Pipeline

The presented pipeline utilizes 2 triplets of calibrated cameras in two rooms separated by a solid, opaque object, in our case a wall. Both triplets of cameras have the same origin point,

and each captured space has dimensions of 4.5x4.5x2 (X, Y, Z) meters. The 3D pose estimation method predicts the 3D poses, which are later passed through a spatial 3D tracker, where the detections are matched using the center of mass from the 3D joints. A person is matched in consecutive detections if the Euclidean distance to the previous center is less than 0.3m. We also utilize a 2D Siamese tracker (Zhu et. al, 2018) onto the projected bounding boxes from one camera to track people of interest, e.g., the AR glasses user, with higher accuracy. The 2D tracker in essence improves the accuracy of the 3D tracker. In Figure 1, we showcase the 3D pose estimation pipeline for the two rooms. As illustrated, an intermediate layer produces a message containing information about the 3D poses, the tracking IDs and affiliations, which is transmitted to the AR devices via a message broker.

With regards to the projection of the 3D poses onto the AR glasses, we utilize an Inertial Measurement Unit (IMU) to determine the orientation of the head. The initial yaw value cannot be extracted from gyroscope and accelerometer sensors, and in this respect, we have experimented with CNNs for head pose estimation, April Tags and Plane Detection. Experimental results indicate superior performance of April Tags.

The initial yaw value and IMU data are used for the head pose initialization. We fuse data from the gyroscope and accelerometer with different weights and the rotation matrix $\boldsymbol{R}$ can be easily computed from the fused yaw, pitch, roll angles. The translation vector $\boldsymbol{t}$, is assumed to be the head position predicted from the 3D pose estimation method. Each AR device has a unique identifier, enabling it to retrieve the correct 3D skeleton from the predicted ones and obtain $\boldsymbol{t}$. The system can handle more than one AR users, and their affiliations are decided by the unique ID of the AR device they are wearing. The 3D poses can then be projected to AR glasses using the camera intrinsic parameters and $\boldsymbol{R}|\boldsymbol{t}$ matrix.

Experimental Results

In this section, a quantitative evaluation of specific components of the system is presented as well as qualitative results for the whole X-Ray pipeline.

Towards evaluating the accuracy of the 3D pose estimation in the wild, we used frame sequences captured by 3 calibrated cameras. We deployed state-of-the-art 2D pose estimation method on the synchronized frames from each camera and deployed triangulation from the OpenCV Structure from Motion (SfM) library, to obtain the ground truth 3D joints location, by exploiting the 2D joints from each view, the camera intrinsic parameters and the $\boldsymbol{R}|\boldsymbol{t}$ matrix. The evaluation was done for 90 frame triplets. As evaluation metric, the Mean Per Joint Positional Error (MPJPE) was calculated:

$$MPJPE = \frac{\sum_i^n \|\hat{y}_i - y_i\|_2}{n} \quad (1)$$

where $\hat{y}_i$ is the i-th joint ground truth, y_i the model prediction and n the number of joints. In our experiment the MPJPE was 83 mm while the same model when evaluated on the Panoptic dataset with 3 cameras has an error of 31 mm. Thus, there is a slight increase in the MPJPE for our setup, however the 8 cm of 3D error is considered acceptable for the presented application. Moreover, in the experiments we have conducted latency is more important than the 3D error and we manage to achieve real-time inference for the whole pipeline (60ms). In Figure 2 we

provide the projection of 3D poses to the 3 cameras in our setup, as well as the visualization of 3D skeleton in space.

In order to measure the efficacy of the head pose estimation algorithms for the initial yaw value, we conducted an experiment, where we used an Intel RealSense camera mounted on a stationary tripod. We considered the wall point where the distance between the wall and the camera was at minimum to be the zero yaw angle in these experiments. We measured the mean absolute error for each Euler angle, in 19 different yaw positions and for each yaw position in 5 different pitch positions. In total 85 positions were measured. Results are presented in **Table 1** and **Table 2**, with the first table referring to the April Tag pose estimation while the second one to Plane Detection. It is evident that pose estimation from April Tag is superior to the head pose prediction from Plane Detection. It should be noted that the roll angle cannot be estimated from the Plane Detection method and thus it is excluded. In Figure 3, the experimental setup is presented. We should mention that April Tag is superior to CNN based methods as well, as they have a mean absolute error (MAE) greater than 3.25 in the datasets they are trained on and 6-10 MAE at best, in more demanding datasets (Zhou et. al, 2023).

Table 1. Camera Pose error per angle April Tag

Angle	Mean	Median	Standard Deviation
Pitch	1.632	1.335	1.333
Yaw	1.568	1.24	1.082
Roll	3.591	2.707	3.171

Table 2. Camera Pose error per angle Plane Detection

Angle	Mean	Median	Standard Deviation
Pitch	19.64	3.807	29.122
Yaw	8.785	3.455	11.134

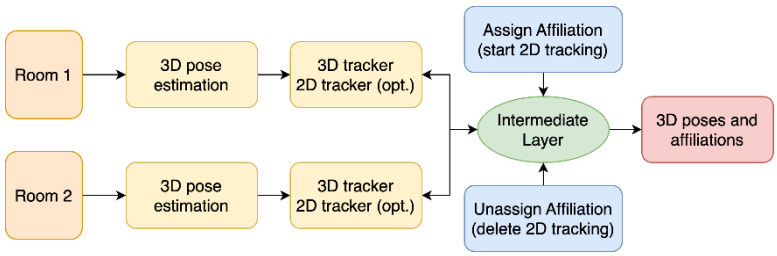

Figure 1. 3D pose estimation and tracking pipeline.

Figure 2. Projection of 3D poses and 3D visualization

Figure 3. Angle exp. setup

Finally, we provide qualitative results for the projection of the 3D poses onto the AR glasses in Figure 4, where the outside space along with the detected person are visualized. In the left image in Figure 4, the red bounding box indicates the person encompassed within is being tracked, while the cyan bounding box is the bounding box from the projected joints. In Figure 5 we provide 2 views with the projected 3D poses back to the cameras while in Figure 6, we provide the AR visualization onto the glasses themselves. The projected poses in Figure 6 are correct and they are in-line with the projected poses in Figure 5. An additional example of AR visualization is provided in Figure 7.

Figure 4: Outside Room

Figure 5: Inside Room

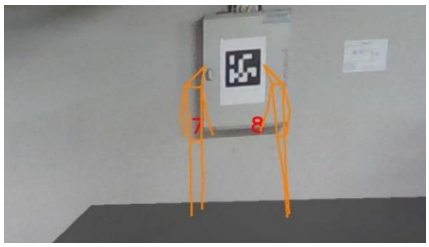

Figure 6: AR Visualization

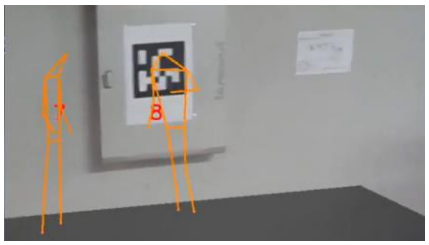

Figure 7: AR Visualization II

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883297 (project DARLENE).

References

Apostolakis, K. C., Dimitriou, N., Margetis, G., Ntoa, S., Tzovaras, D., & Stephanidis, C. (2022). DARLENE–Improving situational awareness of European law enforcement agents through a combination of augmented reality and artificial intelligence solutions. Open Research Europe, 1(87), 87.

Ye, H., Zhu, W., Wang, C., Wu, R., & Wang, Y. (2022, October). Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In European Conference on Computer Vision (pp. 142-159). Cham: Springer Nature Switzerland.

Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., & Hu, W. (2018). Distractor-aware siamese networks for visual object tracking. In Proceedings of the European conference on computer vision (ECCV) (pp. 101-117).

Zhou, H., Jiang, F., & Lu, H. (2023). DirectMHP: Direct 2D Multi-Person Head Pose Estimation with Full-range Angles. CoRR