IoT Camera-Assisted Localization for AR in Simulated and Real-World Environments

Yuhe (Alice) Hu, Ying Chen, Maria Gorlatova
Intelligent Interactive Internet of Things (I^3T) Lab, Duke University, Durham NC

Motivation

Background

- Augmented reality (AR) has the potential to revolutionize industries from industrial to consumer applications
- To achieve broader market appeal, AR must consistently and precisely track user movement and location [1]

Existing AR tracking method

- Widely used first-view AR headset cameras lack global references and accumulate drift due to a limited field of perception [2]

Our AR tracking method

- Third-view IoT cameras provide 'collaborative perception' that can drastically improve the predictions of both user location and pose [3]
- This study uses third-view IoT cameras to evaluate user position detection accuracy in Unity-based simulated 3D space and real-world lab environments.

Experimental Design

Markers for Movement Tracking

Marker 1: AruCo Code

Marker 2: Colored

Marker 3: Unity Logo

- Attached each marker to headset to track user position and orientation
- 3 trials were conducted for each marker, across different marker and camera settings

Data Collection

Simulated Environment

Human follows pre-programmed trajectory

Real-World Environment

Human moves along fixed path in lab

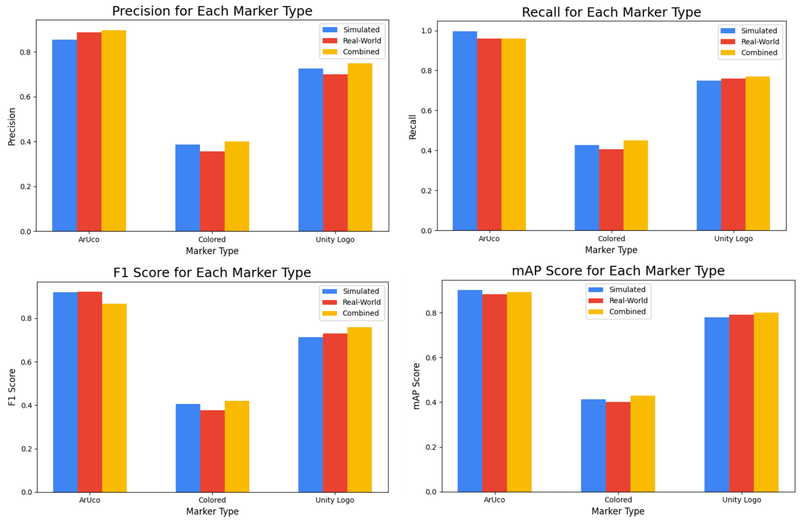
Data Analysis

Deterministic Pattern Recognition (OpenCV Toolbox)	Machine Learning-Based Detection (YOLO v5)
Training Data: 1,600+ images at 640x640 resolution for ArUco marker detection **Processing**: • 30 fps on a CPU at 2.5 GHz **Accuracy Parameters**: • Minimum Margin: 0.25 • Maximum Corner Error: 0.02 (pixels) **Limitation**: Only detects ArUco markers **Avg Reference Time**: 50 ms per marker	**Training Data:** 1,600+ images at 640x640 resolution per model trained **Processing:** • 60 fps on a GPU at 1.59 GHz (T4) • CPU at 2.5 GHz **Accuracy Parameters:** • Confidence Threshold: 0.5 • NMS Threshold: 0.45 **Avg Reference Time:** 30 ms per marker

Analyzing Data with Yolov5 Neural Network

Data Labeling

- 1000+ images were collected for each marker and environment and labeled with a 98.5% inter-annotator agreement score via Roboflow Autodistill
- Labels verified by human annotators to establish reliable ground truth data



Training

- Three Yolov5 machine learning models were trained for each marker type with train-validate-test split
- Metrics evaluated for each model: Precision, Recall, mAP Score, F-1 Score

Key Takeaways

(1) Marker Variations: *ArUco marker performed best*

- ArUco markers record the highest average precision (87.97%), recall (97.28%), F1-score (90.35%) and mAP score (89.20%)
- Colored markers demonstrate the lowest average precision (38.13%), recall (42.76%), F1-score (40.05%) and mAP score (41.50%)

(2) Environmental Variations: *Combined environments performed best*

- Combining simulated and real-world data resulted in a general performance improvement across all three markers
  - ArUco marker (3.10% increase)
  - Colored marker (7.49% increase)
  - Unity Logo marker (5.27% increase)

(3) Model Variations: *Yolov5 more effective at detecting ArUco*

- Yolov5 has a higher marker detection success rate for ArUco Markers:
  - YOLOv5: 96% of frames successfully detected
  - OpenCV: 68% of frames successfully detected

Analyzing Data with OpenCV Toolbox

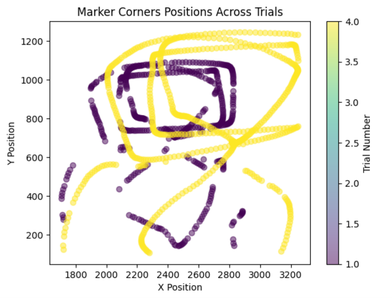
Procedure

- Environmental settings were varied in Unity environment

Trial Number	Marker Size	Camera Position	Camera Angle	Focal Length	Sensor Size
1 (baseline)	*(0.05, 0.05, 0.001)*	*(0.5, 2.5, 0)*	*(90, 0, 0)*	*10*	*36 x 24*
2 (+ marker size)	(0.1, 0.1, 0.001)	(0.5, 2.5, 0)	(90, 0, 0)	10	36 x 24
3 (- marker size)	(0.03, 0.03, 0.001)	(0.5, 2.5, 0)	(90, 0, 0)	10	36 x 24
4 (+ focal length)	*(0.05, 0.05, 0.001)*	*(0.5, 2.5, 0)*	*(90, 0, 0)*	*15*	*36 x 24*
5 (- focal length)	(0.05, 0.05, 0.001)	(0.5, 2.5, 0)	(90, 0, 0)	5	36 x 24
6 (+ sensor size)	(0.05, 0.05, 0.001)	(0.5, 2.5, 0)	(90, 0, 0)	10	22.3 x 14.9
7 (- sensor size)	(0.05, 0.05, 0.001)	(0.5, 2.5, 0)	(90, 0, 0)	10	72 x 48
8 (+ camera angle)	(0.05, 0.05, 0.001)	(0.5, 2.5, 0)	(90, 90, 0)	10	36 x 24
9 (- camera angle)	(0.05, 0.05, 0.001)	(0.5, 2.5, 0)	(90, -90, 0)	10	36 x 24

Key findings:

- Increasing the focal length (x2) led +27.27% frames detected
- Increasing marker size (x2) led to +17.83% frames detected
- Increasing sensor size (x2) led to +8.20% frames detected
- Camera angle adjustments had no effect on frames detected



Conclusions

- Yolov5 Marker Detection performs best when trained in both data from both simulated and real-world environments
- ArUco markers showed the highest average performance while colored markers showed the lowest average performance

Future Work:

- Combining third-view with first-view cameras to further validate the efficacy of the proposed third-view tracking method.
- Expand the dataset with more real-world scenarios and lighting conditions to further test and refine the third-view tracking capabilities
- Integration of more complex AI models beyond YOLOv5

References

[1] Y. Chen, H. Inaltekin, and M. Gorlatova, "AdaptSLAM: Edge-assisted adaptive SLAM with resource constraints via uncertainty minimization," in Proc. IEEE INFOCOM, 2023.
[2] Simon Bultmann, Raphael Memmesheimer, Sven Behnke. (2023). External Camera-based Mobile Robot Pose Estimation for Collaborative Perception with Smart Edge Sensors.
[3] Y. Wen, K. K. Singh, M. Anderson, W. P. Jan and Y. J. Lee, "Seeing the Unseen: Predicting the First-Person Camera Wearer's Location and Pose in Third-Person Scenes," IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021.

Duke UNIVERSITY