

Indoor Receiver Localization using Low-Cost mmWave Wi-Fi Radio and RGB Image Fusion

Genivaldo Charchar¹, João Borges¹, Gabriel Vieira¹, Andrey Silva², Silvia Lins², and Aldebaro Klautau¹

Abstract—Receiver (Rx) localization in indoor scenarios is useful for empowering management pipelines and digital twin systems, among others. For this, many methods exist that leverage radio information such as carrier frequency and/or channel state information (CSI), as well as sophisticated multimodal data, for instance, Light Detection and Ranging (LiDAR), depth cameras and/or radars. However, significant amounts of resources must be invested to obtain this additional information. In this work we leverage easily available Received Signal Strength (RSS) data, alongside RGB camera imagery to provide reasonable estimations of a Rx localization inside an in-laboratory indoor scenario setup.

Keywords—Computer vision, localization, mmWave, multimodal data, Wi-Fi.

I. INTRODUCTION

The capabilities of Wi-Fi [1] and 5G networks [2] allows for exploring indoor positioning use cases, specially given the high directionality of millimeter wave (mmWave) propagation. Also, this radio-based position estimation can be further improved with the fusion of multimodal information such as images [3]. However, many works rely on sophisticated hardware such as radar and special cameras, as well as the availability of communication data e.g. the Channel State Information (CSI), which is not always the case in off-the-shelf and low-cost equipments, increasing the operational cost of such systems. For instance, in [3], Zhang et al. use specialized mmWave radar to provide localization, and enhances it with synthetic RF data obtained from object shapes detected via camera images and computer vision. Bai et al. [1] also leverage Received Signal Strength (RSS) for localization, but uses only traditional Wi-Fi (2.4 GHz/5 GHz), and enhances it with images taken from the receiver (Rx) side with a more sophisticated camera (Kinect v2). In [4], Liu et al. investigate vehicle tracking in adverse conditions by fusing mmWave radar with bounding boxes from object detection. Their study uses specialized radar hardware and simulated data obtained from Simulation of Urban Mobility (SUMO) and Car Learning to Act (CARLA) simulators and no physical measurements. In [5], Sunami et al. focus on dynamic blockage events and use the lower frequency of 5.22 GHz in their experiments, while we use 60 GHz, therefore leveraging much narrower beams. Saikko et al. [6] uses carrier phase and Doppler measurements, alongside Kalman and particle filtering techniques to precisely

locate Rx 3D position in numerical simulations on scenarios evaluated up to 28 GHz. Huan et al. [7] matches 60 GHz CSI fingerprinting with RGB images, both from the synthetic ViWi dataset [8] and provides experiments for outdoor vehicle positioning. Finally, in the work of Mihara et al. [9], Received Signal Strength Indicator (RSSI) is used alongside stereo camera images to provide User Equipment (UE) positioning for blockage prediction, therefore relying on sophisticated imaging hardware. Differently from the previous approaches, in our work we do not use simulated data or specialized radar hardware for localization, instead we rely on radio and image data obtained from off-the-shelf hardware, real measured data, and simple beam sweeping for localization. In this context, this work main contributions are:

- 1) to provide an investigation of low-cost hardware for Rx localization, for both radio and image;
- 2) to do so while considering mmWave radio data;
- 3) to explore the contribution in the position estimation task of RGB image data from the infrastructure side, i.e. a camera placed on the transmitter (Tx) instead of the Rx's;
- 4) finally, to compare the method using RGB image fusion with two different tiers of the implementation, acting as baselines. In other words, the method that uses just the direction estimated from the strongest beam, and also the method that uses the median of the estimate provided by subgroups of beams from the TOP- K ;

II. SYSTEM MODEL

This section presents the proposed Rx localization framework, which combines geometric prior information with radio and image measurements. Subsection II-A describes the Multiple-Inputs Single-Output (MISO) setup and beam sweeping procedure. Subsection II-B presents the beam-to-angle mapping, while Subsection II-C introduces the image-based angular modeling. Subsection II-D details the geometric position estimation, and Subsection II-E describes the subset-based strategy used to improve robustness.

A. Communication system

The considered setup consists of a fixed Base Station (BS) equipped with a Uniform Planar Array (UPA) and a mobile single-antenna UE, forming a MISO system operating at carrier frequency f_c . The BS employs a UPA with M_t antenna elements (M_y columns and M_z rows). The channel between the BS and the UE at each Rx position is represented by an $M_t \times 1$ complex vector. Beamforming is performed using a codebook $\mathcal{W}$ containing W codewords, where each codeword

¹LASSE - Telecommunications, Automation and Electronics Research and Development Center, Federal University of Pará (UFPA), Belém-PA, Brazil. E-mails: {genivaldo.silva, joao.tavares.borges, gabriel.vieira}@itec.ufpa.br; aldebaro@ufpa.br; ²Innovation Center, Ericsson Telecomunicações Ltda., Brazil. E-mails: {andrey.silva, silvia.lins}@ericsson.com

w steers the transmitted beam toward a specific direction. During operation, the BS performs beam sweeping over all codewords and, for each steering direction, the UE measures the RSSI. The corresponding RSS values in dBm are then obtained through the linear mapping in [10],

$$\text{RSS} = 0.0651 \cdot \text{RSSI} - 74.3875. \quad (1)$$

B. Mapping codewords to main lobe directions

Each codeword $w \in \mathcal{W}$ generates a directional beam pattern at the UPA. Assuming isotropic antenna elements arranged along the y and z axes, the array factor for w is given by [11]

$$AF(\theta, \phi) = \sum_{m_y=1}^{M_y} \sum_{m_z=1}^{M_z} w_{m_y m_z} e^{j(m_y-1)\Psi_y} e^{j(m_z-1)\Psi_z}, \quad (2)$$

$$\begin{aligned} \Psi_y &= kd_y \cos \theta \sin \phi, \\ \Psi_z &= kd_z \sin \theta, \end{aligned}$$

where d_y, d_z are inter-element spacings and θ, ϕ, k are elevation, azimuth and wavenumber, respectively. Due to the symmetry of the UPA pattern, $AF(\theta, \phi)$ is evaluated over a discrete spherical grid with angular resolution a_s , where $\theta \in [-\pi/2, \pi/2)$ and $\phi \in [-\pi/2, \pi/2)$. The main lobe direction of each codeword w is defined as the grid point that maximizes the array factor magnitude

$$(\theta_w, \phi_w) = \arg \max_{\theta_i \in \Theta, \phi_j \in \Phi} |AF_w(\theta_i, \phi_j)|,$$

where $\Theta = \{\theta | \theta = -\frac{\pi}{2} + ka_s, k \in \mathbb{Z}\} \cap [-\frac{\pi}{2}, \frac{\pi}{2})$ and Φ is defined analogously for the azimuth angle. Repeating this procedure for all codewords yields the directional mapping of the codebook $\{(\theta_w^*, \phi_w^*)\}_{w \in \mathcal{W}}$, which associates each steering vector with the center of its main lobe.

C. Image-based angular modeling

As an auxiliary source of information, RGB images captured from the Tx Point of View (POV) are incorporated into the localization framework. A You Only Look Once (YOLO)-based detector is used to identify the Rx in each image, and the center of the predicted bounding box, denoted by (u, v) , is taken as the visual reference point. This pixel is then mapped into local camera angles using the image resolution and the adopted horizontal and vertical fields of view. The conversion follows the standard pinhole-camera model where the pixel coordinates are first transformed into normalized image coordinates, then associated with a viewing ray, and finally converted into elevation and azimuth angles [12]. A pitch correction is applied to account for vertical camera tilt. The resulting angular pair $(\theta_{\text{img}}, \phi_{\text{img}})$ is interpreted as an artificial beam, providing an additional directional observation that can be fused with the radio-based beam estimates.

D. Geometric position estimation

The Rx position is estimated geometrically from the known Tx location, the known Rx height, and the angular information of each selected beam. The problem is formulated as a line-plane intersection, in which the beam direction defines the line and the Rx height defines the horizontal plane $z = z_{rx}$. Beams

that are nearly parallel to this plane or that point in the opposite direction are regarded as inconsistent and can be discarded, thereby reducing the impact of outliers and multipath.

The Cartesian estimate of the Rx position is obtained as

$$\hat{\mathbf{p}}_{rx} = \mathbf{p}_{tx} + t^* \mathbf{u}, \quad (3)$$

where $\mathbf{p}_{tx}$ is the Cartesian position vector of Tx and $\mathbf{u}$ is the unit vector in the direction of the selected beam (parametrized by its elevation and azimuth angles). The scalar t^* that places the intersection on the plane $z = z_{rx}$ is given by

$$t^* = \frac{z_{rx} - z_{tx}}{u_z},$$

which follows directly from the parametric equation of the line, $z_{tx} + t u_z = z_{rx}$.

E. Subset combination strategy

Once the TOP- K set has been defined, consisting of the K beams with the highest power, the formation of subsets of cardinality b , with $b \leq K$, is considered. Let $\mathcal{B}_K$ denote the set containing these K beams. From it, the family of all distinct subsets of size b is defined, denoted by $\mathcal{S}$. Thus, the total number of possible combinations is given by

$$N_{K,b} = \binom{K}{b} = \frac{K!}{b!(K-b)!}. \quad (4)$$

Each subset $S_j \in \mathcal{S}$, with $j \in \{1, 2, \dots, N_{K,b}\}$, induces a partial estimate of the Rx position from the individual estimates associated with the beams that compose it. Let $\hat{\mathbf{p}}_i$ denote the Cartesian estimate produced by the i -th beam belonging to subset S_j . The partial estimate associated with this subset is then defined as

$$\hat{\mathbf{p}}_{S_j} = \frac{1}{b} \sum_{i=1}^b \hat{\mathbf{p}}_i, \quad (5)$$

where b represents the number of elements in each subset. Thus, each subset generates a partial estimate, and the set of all such estimates can be written as

$$\mathcal{P} = \{\hat{\mathbf{p}}_{S_1}, \hat{\mathbf{p}}_{S_2}, \dots, \hat{\mathbf{p}}_{S_{N_{K,b}}}\}.$$

The final position estimate is then obtained through a consensus operation applied over $\mathcal{P}$ (6). In this work, the median is adopted as the aggregation statistic due to its greater robustness in the presence of discrepant partial estimates, which may arise from outliers beams or multipath-related trajectories.

$$\hat{\mathbf{p}}_{\text{final}} = \text{median}(\mathcal{P}) \quad (6)$$

III. SCENARIO AND DATASET DESCRIPTION

The experiments were conducted in an indoor laboratory environment using a fixed BS and a mobile robotic platform acting as the UE. The BS was equipped with a 6×6 UPA, i.e., $M_y = 6$ and $M_z = 6$, with the four corner elements disabled. Each antenna element consisted of a crossed dipole, and the measurements were collected at $f_c = 60.48$ GHz using a codebook whose angular coverage is approximately limited to $[-40^\circ, 40^\circ]$ in both elevation and azimuth.

TABLE I
SYSTEM ARCHITECTURE

Parameter	Value
Operating frequency	60.48 GHz
Tx and Rx height	1.47 m and 0.87 m, respectively
Transmit array (Tx)	6x6 UPA (36 elements, 4 disabled)
Receive antenna (Rx)	Single element
Beamforming method	Beam sweeping
Hardware	MikroTik wAP 60G
Standard	IEEE 802.11ad [13]
Antenna type	Crossed-dipole
Antenna polarization	Dual (Vertical+Horizontal)
Measurement duration	50 s (250 samples) each position
RSS sampling interval	200 ms
Camera type	Skydrone FPV3 (Skylab)
Image resolution	1280x720 pixels (Horizontal x Vertical)
Camera tilt	-20° (Elevation)

The robotic platform, equipped with a mmWave radio with a single active antenna element and an RGB camera, followed a predefined route inside the environment. Along this trajectory, 58 valid Rx positions were considered, each one the platform remained static, and both channel information and the corresponding RGB image data were collected. The system architecture is summarized at Table I. For the localization stage, only the first RSS capture of each position was used [10]. In this sense, a two-dimensional coordinate system was adopted on the room plane. The room dimensions, BS position, trajectory, and Rx positions are shown in Fig. 1.

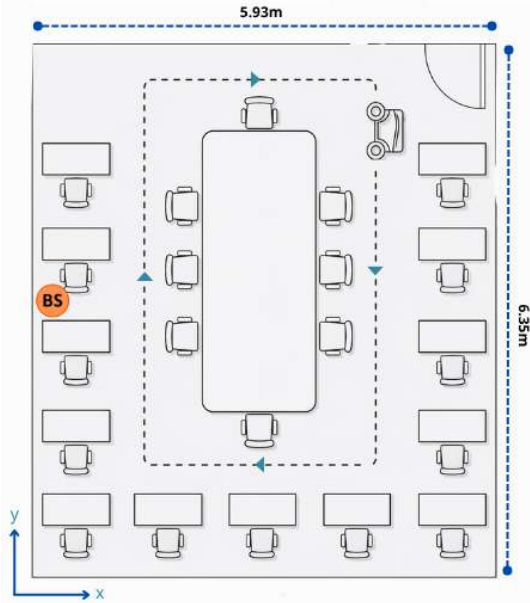


Fig. 1. Experimental indoor scenario, including the room dimensions, adopted coordinate system, BS position, and the trajectory followed by the robotic platform across the Rx positions.

The analyzed positions included both on-grid and off-grid conditions with respect to the codebook angular discretization. In the first case, the Rx direction was aligned with the angular grid and could be more represented by the available beams. In the second, it did not exactly match the discretized directions, resulting in a more challenging off-grid localization condition.

IV. PROPOSED METHOD

The proposed method estimates the Cartesian position of the Rx by combining geometric prior information with radio measurements obtained from mmWave beam sweeping. The method assumes that the Tx position is known, that the Rx lies on a plane with known height, and that each beam index can be associated with an angular direction through the adopted codebook and antenna model. In the image-assisted version, RGB images captured from the Tx POV are also used as an additional source of angular information.

The process starts from the RSSI values associated with all beams at a given Rx position. The TOP- K beams with the highest received power are selected under the assumption that, in line-of-sight conditions, stronger beams are more likely to point toward the true Rx direction. Each selected beam is then mapped to its corresponding angular pair (θ, ϕ) , representing elevation and azimuth. This association is obtained from the beam radiation pattern, where the beam direction is defined as the angular location of maximum transmitted power under an isotropic antenna element approximation. In the image-assisted case, the receiver is detected in the RGB image and the center of the predicted bounding box is mapped into a local angular pair $(\theta_{\text{img}}, \phi_{\text{img}})$. This direction is interpreted as an artificial beam and added to the set of radio-based angular observations.

Once the angular information is available, each radio beam, and, when applicable, the image-derived artificial beam, is converted into a 3D unit direction vector and projected onto the Rx plane, yielding individual Cartesian estimates. Rather than relying on a single beam, the method explores all subsets of cardinality b formed from the TOP- K set, according to (4). In the image-assisted version, the image-derived estimate is also included in the subset formation. Then, each subset produces a partial position estimate through the arithmetic mean of the Cartesian estimates of its elements. Finally, all partial estimates are aggregated through the coordinate-wise median, as given in (6). Therefore, the image-derived estimate contributes both to the partial subset estimates and to the final consensus-based localization result.

V. RESULTS AND DISCUSSIONS

This section presents the localization results obtained for the 58 Rx positions in the dataset. All evaluated strategies are based on the geometric beam-wise estimation described in the section IV and differ in how directional information is selected and aggregated. Three approaches were considered, the first is a TOP-1 baseline based on the strongest beam only, the second is a subset-based strategy using the TOP- K beams with median-based consensus, and the last is an image-assisted version in which RGB-derived angular information is incorporated as an additional information. Unless otherwise stated, the subset-based experiments were performed with TOP- $K = 5$ and subset cardinality $b = 4$, since this configuration provided the best trade-off between accuracy and robustness.

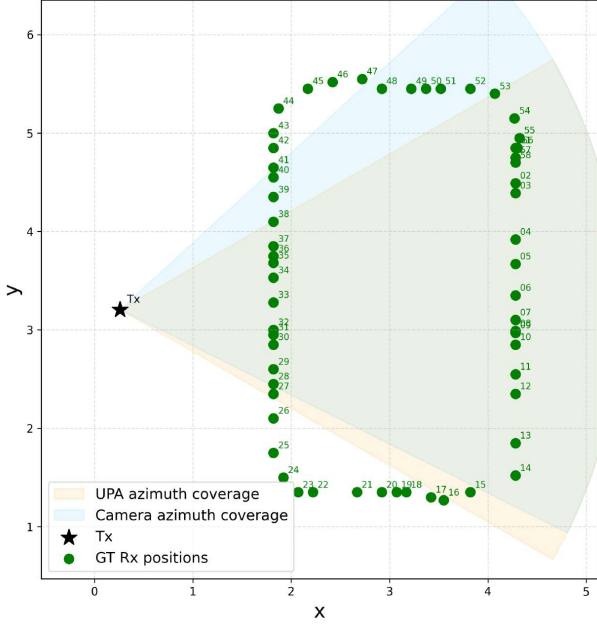


Fig. 2. Ground-truth Rx positions in the scenario, together with the Tx location and the azimuthal coverage regions of the UPA and the RGB camera.

TABLE II

2D LOCALIZATION PERFORMANCE FOR THE EVALUATED METHODS

Method	Mean [m]	Variance [m ²]	Std. Dev. [m]
Top-1	1.1394	0.4933	0.7024
Subset	0.7025	0.1774	0.4212
Subset+image	0.6374	0.1861	0.4314

A. Experimental protocol and evaluation metrics

The performance was assessed through the Cartesian coordinate errors and the planar localization error, defined as

$$e_{2D} = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2}, \quad (7)$$

where (x, y) and $(\hat{x}, \hat{y})$ denote the true and estimated Rx positions, respectively. These metrics allow evaluating both coordinate-wise behavior and overall localization accuracy. In addition, spatial error plots were used for qualitative analysis, while the main statistical indicators derived from the 2D localization error are summarized in Table II.

B. TOP-1 baseline

The first evaluated strategy uses only the strongest beam, such that the Rx position is estimated directly from the angular information associated with a single directional observation. This configuration serves as a baseline, since it represents the most direct use of the proposed geometric formulation without any additional combination or consensus mechanism.

Fig. 3 shows, in blue, the x - and y -component localization errors for the TOP-1 configuration. Although valid estimates are obtained for all positions, the largest errors occur mainly in positions 1-16 and 50-58 for the x -component, and around positions 16, 44-46, and 54 for the y -component as illustrated in Fig. 2. These regions are more sensitive to angular mismatch, especially when the receiver is farther from the Tx or when its direction is not well represented by a single codebook beam.

As a result, the finite angular discretization of the codebook directly translates into larger Cartesian localization errors.

C. Subset-based beam localization

The second strategy explores the angular diversity of the TOP- K beams. As shown in Fig. 3, the subset-based approach yields a more balanced error distribution than TOP-1, reducing the mean 2D error from 1.1394 m to 0.7025 m, i.e., an improvement of 38.34%. This gain results from combining multiple strong beams, which reduces the dependence on a single steering direction and partially mitigates the codebook angular discretization. In addition, the median-based consensus step emphasizes geometrically consistent subset estimates while reducing the effect of noisy or outlier solutions.

It is also worth noting that the spherical-to-Cartesian mapping induces a non-uniform spatial distribution of the estimated directions. As a result, the angular grid becomes effectively denser near the Tx and sparser at larger distances. This helps explain why the subset strategy is particularly beneficial at farther receiver positions, where single-beam estimation is more strongly affected by discretization effects. Since each subset estimate is obtained by averaging beam-wise projections, the final solution may become slightly biased toward the central region of the mapped area. Moreover, the formulation assumes dominant Line-of-Sight (LOS) behavior; under Non-Line-of-Sight (NLOS) conditions, reflected paths may shift the apparent beam directions and consequently bias the estimated position.

D. Image-assisted subset localization

The last evaluated strategy extends the subset-based beam localization method by incorporating image-derived angular information. More specifically, the radio location in the RGB image is used to generate an additional directional observation from the Tx POV. This visual contribution is then combined with the beam-based estimates in order to further constrain the Rx position, especially under off-grid conditions.

Fig. 3 shows that the image-assisted strategy leads to a more concentrated error profile than the beam-only subset approach. This improvement is also confirmed by the quantitative results: the mean 2D error is further reduced to 0.6374 m, corresponding to an improvement of approximately 44.06% with respect to the TOP-1 baseline. The benefit of the image-derived contribution is not uniform along the trajectory. It is more evident in positions 4-14 and 48-51, where the image-assisted solution reduces the x - or y -component errors more consistently than the beam-only subset method. These regions correspond to outer portions of the covered area, where the Rx is farther from the Tx or closer to the limits of the angular coverage. In such cases, the visual observation acts as a complementary artificial beam, refining the directional inference and making the final localization less sensitive to codebook discretization and angular resolution effects.

This gain was obtained even in the presence of practical limitations affecting the image-based angular modeling, including distortions that become more pronounced away from the image center, the adoption of a simplified pinhole camera

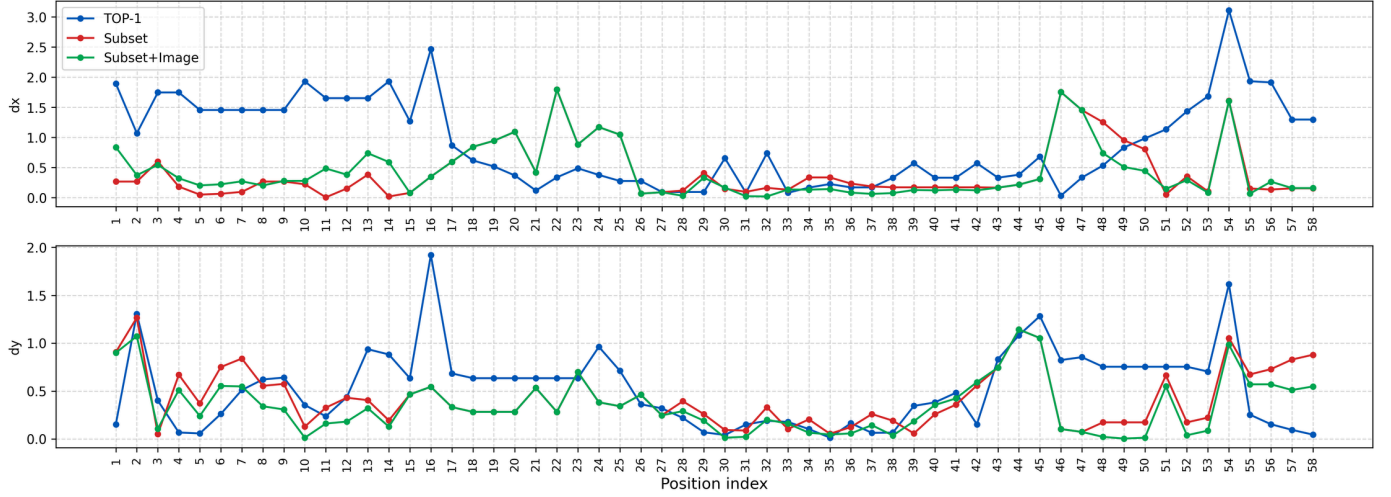


Fig. 3. Comparison of the Cartesian localization errors in the x - and y -components for the TOP-1, subset-based, and image-assisted subset-based methods.

model, and possible vertical and horizontal camera tilts, all of which introduce uncertainty in the pixel-to-angle mapping process. Even so, the visual contribution remained sufficiently informative to improve the geometric consistency of the final estimate and to complement the radio-based subset strategy.

VI. CONCLUSIONS

In this work we used a low-cost mmWave setup alongside an RGB camera to achieve multimodal radio and image fusion based indoor Rx positioning estimation. We leveraged easily accessible radio information via RSS data gathering from the Tx beam sweep over a codebook composed by 64 directional beams. These beams orientations, alongside the prior information of the Rx starting position, allowed for the system to obtain a rough estimate of the Rx localization per-beam. This value was further refined using subgroup of beams, from which first were extracted the mean for the Cartesian coordinates and associated with each subgroup, then from the median of these subgroup position estimates we obtained a refined estimate. Finally, we used a object detection approach for another estimation refining by leveraging YOLO bounding boxes, which were placed over the detected Rx on RGB images taken from the Tx POV. This contribution added on the "subgroup" approach provided the final estimate, which gives roughly an position estimation error of less than 1 meter for the studied case. Future works include using a more accurate crossed-dipole antenna model for improved beam characterization, using improved camera models to reduce pixel-to-angle mapping errors, and developing adaptive strategies to choose parameters such as TOP- K , subset cardinality, and the use of image information according to the confidence of the individual radio and image-based estimates. In addition, more challenging scenarios with multiple Rx's, obstacles and NLOS conditions can be considered.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001; the Conselho Nacional de

Desenvolvimento Científico e Tecnológico (CNPq); the Brasil 6G project (01245.020548/2021-07), supported by RNP and MCTI; the Innovation Center, Ericsson Telecomunicações Ltda., Brazil; OpenRAN Brazil - Phase 2 project (MCTI grant N° A01245.014203/2021-14); Project Smart 5G Core And MultiRAN Integration SAMURAI (MCTIC/CGI.br/FAPESP under Grant 2020/05127-2).

REFERENCES

- [1] N. Bai, G. Chen, R. Hou, and F. Ying, "A novel WiFi signal and RGB image fusion positioning method for manufacturing workshop," 2020.
- [2] M. Abuyaghi, S. Si-Mohammed, G. Shaker, and C. Rosenberg, "Positioning in 5G Networks: Emerging Techniques, Use Cases, and Challenges," *IEEE Internet of Things Journal*, vol. 12, no. 2, pp. 1408–1427, 2025.
- [3] D. Zhang, A. Prabhakara, S. Munir, A. C. Sankaranarayanan, and S. Kumar, "A Hybrid mmWave and Camera System for Long-Range Depth Imaging," 2021.
- [4] H. Liu, J. Wan, P. Zhou, S. Ding, and W. Huang, "Augmented Millimeter Wave Radar and Vision Fusion Simulator for Roadside Perception," *Electronics*, vol. 13, no. 14, 2024.
- [5] T. Sunami, S. Itahara, Y. Koda, T. Nishio, and K. Yamamoto, "Computer vision-assisted single-antenna and single-anchor RSSI localization harnessing dynamic blockage events," *arXiv preprint arXiv:2107.04770*, 2021.
- [6] A. Saikko, J. Talvitie, and M. Valkama, "High-precision 3D location and orientation tracking: quaternion-based approach using cellular carrier phase and Doppler measurements," *Journal on Advances in Signal Processing*, 2026.
- [7] O. Huan, T. Luo, and M. Chen, "Multi-Modal Image and Radio Frequency Fusion for Optimizing Vehicle Positioning," *IEEE Transactions on Mobile Computing*, vol. 24, no. 2, pp. 696–708, 2025.
- [8] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A Deep Learning Dataset Framework for Vision-Aided Wireless Communications," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1–5, 2020.
- [9] S. Mihara, S. Ito, T. Murakami, and H. Shinbo, "Positioning for User Equipment of a mmWave System Using RSSI and Stereo Camera Images," 2021.
- [10] V. Conceição, F. Nunes, J. Borges, C. Nahum, I. Correa, S. Lins, and A. Klautau, "Low-Cost Experimental Platform for AI-Based Beam Management Using WiFi Radios," 01 2025.
- [11] C. Balanis, *Antenna Theory: Analysis and Design*. Wiley, 2015.
- [12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, pp. 153–156. Cambridge University Press, 2 ed., 2004.
- [13] R. S. I.-R. International Telecommunication Union, "IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band," Standard IEEE Std 802.11ad-2012, IEEE, 2012.