Egocentric Room Location Classification using Deep Neural Network Measuring Uncertainty

Miran Seo, Sangjoon Park, Soyeon Lee, Blagovest Iordanov Vladimirov, and Joo Dong Yun*
Digital Convergence Research Laboratory,
Electronics and Telecommunications Research Institute,
Daejeon, Republic of Korea
seomr8@gmail.com, {sangjoon, sylee, vladimirov, joodong1}@etri.re.kr

***Abstract*—The rise of mobile applications in Virtual Reality (VR) and Augmented Reality (AR), particularly those using Head-Mounted Displays (HMDs), underscores the need to understand egocentric perspectives. This paper addresses the room-level localization challenge—identifying the room a user is in from an egocentric image—by framing it as a classification problem with a deep neural network. While deep learning has achieved remarkable success in conventional image classification, room classification from egocentric images introduces unique challenges due to variability and ambiguity in the user's perspective. Unlike typical datasets that provide clear visual data, egocentric views often lack sufficient detail, making uncertainty estimation crucial for achieving accurate results. Our approach not only advances egocentric localization but also holds potential for improving navigation and context-aware applications in AR/VR environments. We propose a novel strategy for uncertainty estimation and validate it with a custom dataset. Experimental results reveal significant performance improvements, achieving near-perfect accuracy by effectively managing ambiguous samples.**

***Index Terms*—neural network, room classification, uncertainty estimation, indoor localization.**

I. Introduction

Virtual Reality (VR) and Augmented Reality (AR) are becoming integral to daily life, driving research into mobile applications, particularly in indoor localization. This rising interest has intensified the focus on understanding egocentric images [1] captured from the user's perspective through Head-Mounted Displays (HMDs). Unlike outdoor systems that must adapt to varying light and seasonal conditions, indoor localization primarily deals with challenges such as the lack of satellite-based positioning. To overcome these issues, researchers are investigating alternative methods that rely on additional signals [2], [3], [4].

In this paper, we advance indoor localization for large, multi-room areas by focusing on room location classification, an underexplored approach in this field. We utilize a newly generated dataset designed specifically for this purpose and rely exclusively on egocentric images for classification, while avoiding additional signals. This approach provides a fresh perspective on accurately identifying room locations within indoor environments.

*Corresponding author: joodong1@etri.re.kr

II. Proposed Approach

Consider a classification dataset $\{(x_i, y_i)\}_{i=1}^{N}$, which consists of N pairs, where x_i represents an image with a specified number of pixels, and y_i is the corresponding one-hot encoded label with n_c classes. Training a classification neural network f involves solving the following optimization problem, which minimizes the cross-entropy loss by adjusting the model parameters Θ:

$$\min_{\Theta} -\sum_{i} \langle y_i, \log(\sigma(f(x_i;\Theta)))\rangle, \tag{1}$$

where $\log$ operates pointwise, and σ denotes the softmax function, which transforms the neural network's output into a probability vector, ensuring that the sum of its elements equals one, with each element lying between 0 and 1. Let us define the probability vector as follows:

$$p_i = \sigma(f(x_i;\Theta)) \tag{2}$$

In transfer learning, models are typically initialized with pretrained weights from large public datasets like ImageNet and then fine-tuned for fewer epochs. For our model f, we use a pretrained ResNet-18 [5].

A. Measuring Uncertainty

For a given image x_i, we define a measure of uncertainty τ in a classification task using entropy, which is calculated as follows:

$$\tau(x_i) = -\frac{1}{\log n_c}\langle p_i, \log(p_i)\rangle, \tag{3}$$

where p_i is obtained from (2), and $\log n_c$ acts as a normalizing factor, reflecting that maximum entropy occurs when all components are equal to $1/n_c$. This measure differs from the cross-entropy loss in (1), which is utilized during the training phase; instead, (3) is computed once per sample during the inference phase. Since the softmax function output can be viewed as a probability distribution, the equation above calculates the entropy of the predicted probabilities from the neural network. With Θ fixed, this measure can be used as a criterion to eliminate uncertain samples.

Given a threshold η, an image x_i is classified only if $\tau(x_i) \leq \eta$; otherwise, it is excluded from classification. The effectiveness of this uncertainty measure τ will be empirically validated in the experiments discussed later.

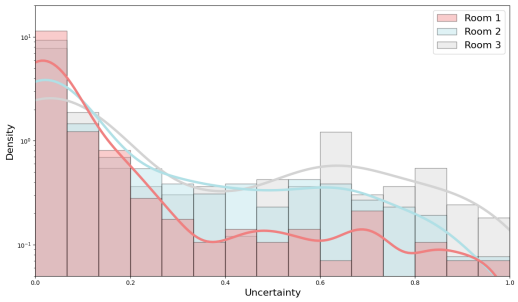

Fig. 1. Distribution of uncertainty τ for each room on the dataset.

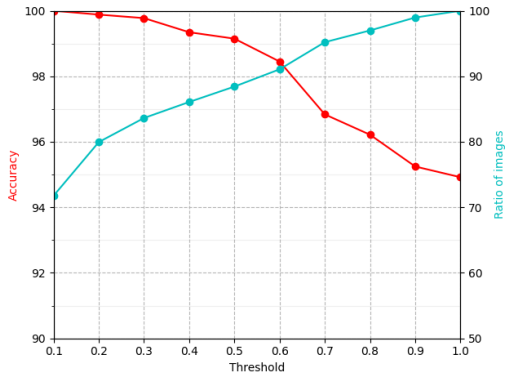

Fig. 2. Accuracy and ratio of non-excluded images applying threshold η.

B. Dataset Description

In this section, we describe a dataset created for room localization, featuring egocentric images from three conference rooms: Room 1, Room 2, and Room 3. While the rooms have similar layouts, Room 1 is distinguished by its square desk arrangement, whereas Rooms 2 and 3 have centrally connected desks and differ in size. This variation provides a basis for testing room classification effectiveness.

To account for transient objects and variations in illumination, images were captured over several days at random times, covering both day and night. The dataset was designed to be diverse, with almost equal proportions of easy and challenging samples. This balance was achieved by including images from both distant viewpoints that capture most of the room and closer viewpoints that provide fewer visual cues. The distribution of images across the rooms is detailed in Table I.

III. Numerical Results

Fig. 1 shows the distribution of the estimated uncertainty, τ, as calculated by (3) for each room. It is important to note that the plot is presented on a semilogarithmic scale, with the y-axis in a logarithmic format. The results indicate that, across all rooms, the majority of samples exhibit low uncertainty. However, the uncertainty distribution for Room 2 and Room 3 is comparatively broader, with a greater number of samples in the high uncertainty region. This suggests that a higher proportion of images from these rooms are more challenging to classify accurately. This observation aligns with expectations, as Room 1 is the most visually distinct in terms of interior design.

TABLE I
Number of Images in Train and Test Sets for Each Room

	Room 1	Room 2	Room 3	Total
Train Set	527	561	396	1484
Test Set	426	390	247	1063

In Fig. 2, our baseline model—using the original ResNet architecture without thresholding by η—corresponds to the rightmost case with $\eta = 1$, achieving an accuracy of 95%. However, by applying a smaller threshold, we observe a significant improvement in accuracy, approaching nearly 100%. Remarkably, at $\eta = 0.5$, the accuracy exceeds 99%. This enhancement is particularly impressive considering that only about 10% of the images are eliminated, allowing us to achieve such high accuracy with the remaining 90% of the dataset. This demonstrates the effectiveness of our method in significantly boosting classification performance.

IV. Conclusion

In this paper, we present a novel approach to indoor localization, focusing on room classification using egocentric images. We generated a custom dataset tailored for this task, enabling us to implement an effective method for estimating and managing uncertainty, which is computed using the entropy of predicted probabilities. By applying an uncertainty threshold, our approach significantly enhanced classification accuracy while discarding only a small portion of ambiguous data. Our method successfully accounted for subtle design variations between rooms, aligning with both intuitive expectations and the observed uncertainty distribution. These results highlight the robustness of our approach in complex indoor environments and its efficiency in handling uncertainty.

Acknowledgment

This work was supported and funded by the ETRI Research and Development Support Program of MSIT/IITP, Republic of Korea. [Project Title: Development of Beyond X-verse Core Technology for Hyper-realistic Interactions by Synchronizing the Real World and Virtual Space/Project Number: RS-2023-00216821]

References

[1] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. M. Farinella, D. Damen, and T. Tommasi, "An outlook into the future of egocentric vision," *International Journal of Computer Vision*, pp. 1–57, 2024.

[2] F. S. Daniş, A. T. Naskali, A. T. Cemgil, and C. Ersoy, "An indoor localization dataset and data collection framework with high precision position annotation," *Pervasive and Mobile Computing*, vol. 81, p. 101554, 04 2022.

[3] X. Zhao and J. Lin, "Theoretical limits analysis of indoor positioning system using visible light and image sensor," *ETRI Journal*, vol. 38, no. 3, pp. 560–567, 2016.

[4] B. Canovas, A. Nègre, and M. Rombaut, "Onboard dynamic rgb-d simultaneous localization and mapping for mobile robot navigation," *ETRI Journal*, vol. 43, no. 4, pp. 617–629, 2021.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.