Original Manuscript ID: sensors-2515264

Original Article Title: "Deep Learning for Counting People from UWB Channel Impulse Response Signals"

To: Sensors Editor

Re: Response to reviewers

Dear Editor,

Thank you for allowing a resubmission of our manuscript, with an opportunity to address the reviewers' comments.

We are uploading (a) our point-by-point response to the comments (below) (response to reviewers), (b) an updated manuscript with yellow highlighting indicating changes (*Supplementary Material for Review*), and (c) a clean updated manuscript without highlights (*Main Manuscript).*

Best regards,

Gun Lee et al.

Reviewer#1, Concern # 1: The word described in this paper, which is experimental in nature, seems to be a fairly limited case (small room and 0 to 5 occupants) and does not address all modalities that were mentioned of interest, for example, hidden and partially hidden people. In addition to these, what happens when objects in the room other than people move around?

Author response: We appreciate the reviewer's constructive observation. We note that for cases where there is motion among the people present, some occlusion of people occur, which is included in our dataset. If the reviewer is referring to people hiding behind objects, that would be out of the scope of our assumption of people being in the line of sight of the Tx and Rx.

Author action: To address the question of object movement, as well as other changes in the environment, we provide example signals of various settings, and add an additional section (Section 8. Discussion), to describe that our method may require training for each different setting.

Reviewer#1, Concern # 2: The paper should provide better motivation for this small case (what is the concern: is the room occupied, is it close to the occupancy limit, etc.) and how it could be expanded to the more interesting cases of large rooms with many people (like conference rooms) and work spaces.

Author response: We would like to thank the reviewer for the constructive suggestion. We agree with the reviewer's comments and have thus addressed the issues in the manuscript.

Author action: We have added additional motivations for the small case, in Section 1. Introduction, as "In this work, we provide experimentation on a small number of people, which may be applied to cases such as measuring elevator capacity." And we have included the following statements in Section 8. Discussion: "We acknowledge that the experiments conducted in this paper are limited in the number of counted people. For applications such as maintaining room occupancy under some limit, it is likely that the number of people will be much higher. In future works, we plan to expand our system for estimation of larger range of people, including experimentation on systems including multiple Rx and Txs."

Reviewer#1, Concern # 3: The method is not compared with other counting methods, such as optical and infrared (IR), nor are the respective accuracies discussed.

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We have incorporated a discussion comparing our method with optical and infrared (IR) counting techniques, including their respective accuracies, into the manuscript. We have included the following descriptions in Section 1. Introduction: "Imaging technology has been traditionally used to determine the number of people indoors, but it can be challenging to accurately count the number of people in poorly lit areas or in places where people are partially obstructed by objects or other people. Moreover, these technologies carry inherent risks, including potential infringements on personal privacy and possibilities for unintentional disclosure of personal information."

Reviewer#1, Concern # 4: The conclusion is important that both model ensembles and signal ensembles is highly effective. But, it seems signal ensembles provides the bulk of the effectiveness. Pleases stat what increase in accuracy is provided by each and then together.

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: In response to this, we have included a detailed description quantifying the individual contributions of the model ensembles and signal ensembles to accuracy. We have included following descriptions in Section 7.5.1~2 : "The results underscore the benefits of model ensembling, particularly its

capacity to augment performance beyond that of individual models in complex and noisy environments. While model ensembling enhances performance without extending processing time, as the models operate separately, it does increase the memory usage depending on the number of models utilized."

"Our experimental, as detailed in Table \ref{tab:multi_signals}, results demonstrate that ensembling multiple signals simultaneously yields a substantial performance improvement. Specifically, when comparing the results of using two signals to using just a single signal, we observe a notable increase of approximately 8\% in performance. As the number of combined signals increases, the system has the potential to reach a 100\% accuracy score. However, a crucial factor to consider is the associated increase in processing time. While the memory usage remains unaffected, the processing time increases proportionally with the number of accumulated signals."

Reviewer#1, Concern # 5: Perhaps the highly computationally expensive model ensembles could be relaxed. How will computation expense scale with larger rooms and more people? Some particular elements that could us clarification.

Author response: We appreciate the reviewer's comments on this matter. Computational complexity depends on the network structures, and the computation will be the same if we apply the same network, regardless of larger rooms or more people. However, we acknowledge that accuracy might degrade with larger rooms or a greater number of people, requiring more Rx and Tx or more CIR ensembles, resulting in increased computations. This is a problem that warrants further research, and thus we have described our plans to investigate this issue in future works, in Section 8. Discussion, added in this revision.

Author action: None

Reviewer#1, Concern #6: Figure 1: please provide dimensions for room size and antenna spacing. Figure 2 and 3: it is not clear if these are single CIRs or averaged/collections

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We add descriptions about environment in Section 3.1. UWB Data Acquisition:

"Room dimensions were 12m x 8m x 2.6m. In a Line of Sight (LoS) environment, the two modules were placed 4m apart in a straight line, each positioned 1m above the floor and fixed using a tripod."

We add caption about fig 2,3. : "Each plot represents the mean and variation calculated from 500 individual CIR waveforms."

Reviewer#1, Concern # 7: E.G. Table 3 and others: When the term accuracy is used, is this a simple binary value (count correct or not)? It could perhaps be interesting to stat how far off the count was (say returned a value of 4 when 5 was expected, wish is less bad than returning a value of 1 when 5 was expected).

Author response: In our study, 'accuracy' indeed refers to a simple binary value, indicating whether the count is correct or not. We understand the interest in evaluating the degree of error; therefore, we present Table 7, which shows the error rate calculated using the Mean Squared Error (MSE) loss term. This metric provides an understanding of which errors are more significant. However, to keep our performance measures straightforward and comprehensible, we mainly report accuracy in our study.

Author action: None

Reviewer#1, Concern # 8: Overall, the paper is reasonably well written. However, there are several places with sentence fragements, i.e., lase sentence of abstract (was there more needed?), lines 44-45, lines 275-276, etc.

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We modified the sentence :

"We demonstrate that deep neural networks can accurately classify the number of people within the Line of Sight (LoS), achieving an 99\% performance and efficiency with respect to both memory size and FLOPs (Floating Point Operations Per Second). "

"In recent works~\cite{mohammadmoradi2017room,de2021cir}, methods have been developed to estimate the number of people in a room based on the waveform of the CIR from UWB transceivers."

"We conducted our work in a Linux OS, which served as the fundamental programming environment, using Python version 3.6.9."

Reviewer#2, Concern # 1: You have failed to explain why using Deep Learning and the large number of FLOPs is necessary over simple methods that can look at the variance like SVD, or PCA. This paper would be vastly improved if you presented simple methods and the amount of FLOPs for comparison against your method.

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: In response to this, we have included a new section 7.2 in the paper where we present baseline comparisons using raw data input and PCA. In section 7.3, we demonstrate how our neural network method performs better than these baseline methods. Furthermore, we also compare our method with the SVD approach in section 7.5.2, where we deal with multiple CIRs. This way, we offer a comprehensive comparison of the computational costs and performance across different methodologies.

Reviewer#2, Concern # 2: You need to explain why there is more variance at the higher index numbers?

Can you use the variance for the upper index values (30~64) to count the number of people?

You need to explain why you are performing machine learning and neural networks for this system over the simpler methods such as SVD or PCA, The variance signals are clearly larger for the upper index values (30~64).

There also appears to be a difference in the signals in the lower index numbers (1~30) that could indicate whether the people are moving or not.

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We add descriptions about that why It makes a difference between lower index and higher index :

"Signals with a higher index tend to show more variation than those with a lower index. This is primarily due to the nature of signal wave propagation. Lower index signals are predominantly formed by direct waves, which have a straightforward path. In contrast, higher index signals are largely formed by indirect waves that reflect off walls or objects, causing a greater degree of variation."

Reviewer#2, Concern # 3: Explain why are the percentages larger than the case for 3 or 4 people?

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We add descriptions about that why 3 or 4 people performance better than 5 people:

"Table \ref{tab4} shows the average estimation accuracy for different numbers of people. Interestingly, the accuracy for 3 and 4 people is lower than for 5. This suggests that while the waveforms for 3 and 4 people are complex, they are more easily confused with each other, leading to decreased accuracy."

Reviewer#2, Concern # 4: When did you minimize the FLOPs? You consistently chose the method with small percentage increases that also had lare FLOP increases?

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We modified the sentence :

"Our experiments focused on achieving the maximum possible accuracy in estimating the number of people, while concurrently enhancing the efficiency of model parameters and FLOPs."

Reviewer#2, Concern # 5: If 99% is the goal, then this should have been stated in the beginning..

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We modified the sentence :

"An ensemble of two models and four CIR signals, as the best tradeoff between performance and efficiency, achieved 99\% accuracy in 321,000k FLOPs with 6,566k parameters."

Reviewer#2, Concern # 6: Since UWB is not used very much, then it would be unlikely that there would be any other UWB jammers present. This would not be the case for wifi and bluetooth, which the people you are counting may bring into the room.

Author response: We assume that UWB sensors are already installed in the environment and that individuals don't need to carry UWB jammers. When people enter the area within the antenna's range, our method can then detect and count the number of individuals present.

Author action: None

Reviewer#2, Concern # 7: Label the transmitter, receiver, room size. the area where the people will be that in considered to be LoS. Is the entire room considered LoS?.

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We modified figure which denote Tx & Rx and add the description:

"Room dimensions were 12m x 8m x 2.6m. In a Line of Sight (LoS) environment, the two modules were placed 4m apart in a straight line, each positioned 1m above the floor and fixed using a tripod."

Reviewer#2, Concern # 8: You have failed to explain why using Deep Learning and the large number of FLOPs is necessary over simple methods that can look at the variance like SVD, or PCA. This paper would be vastly improved if you presented simple methods and the amount of FLOPs for comparison against your method.

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We have included the section 7.2 about PCA method and add the descriptions in Section 7.5.2 about SVD method. We can show that our experiment has better performance. :

"We provide experimental evaluations of three basic classification methods, decision tree classifier, Naive Bayes classifier, and k-nearest neighbor (K-NN) classifier. We apply train these classifiers on the raw CIR data vector, and the lower dimensional vector constructed by applying Principal Component Analysis (PCA) reconstruction with only a portion of the principal components, as done in the method for face recognition~\cite{turk1991eigenfaces}.

With a single signal, the performance appears to be similar between the PCA method and raw data, as shown in Table~\ref{tab_pca}. However, the PCA method helps to reduce the dimensionality of the data, which decreases the computational cost of training.

As for the K-NN classifier, it achieved an accuracy of approximately 66\%. Although this can be improved with multiple iterations and adjustments to the components, the current level of accuracy is not strongly indicative of a good performance."

"Prior research \cite{de2021cir, han2022hrp} indicates that accumulating 100 CIRs and performing SVD can achieve excellent accuracy. However, this approach requires at least 30 seconds to accumulate the CIRs and carry out the SVD. While we followed the descriptions of prior research, using the three features of 2nd slope, average slope and area under curve as features, for training the classifiers, we observed that it underperformed the neural network based classifiers.

Our experimental, as detailed in Table \ref{tab:multi_signals}, results demonstrate that ensembling multiple signals simultaneously yields a substantial performance improvement. Specifically, when comparing the results of using two signals to using just a single signal, we observe a notable increase of approximately 8\% in performance. As the number of combined signals increases, the system has the potential to reach a 100\% accuracy score. However, a crucial factor to consider is the associated increase in processing time. While the memory usage remains unaffected, the processing time increases proportionally with the number of accumulated signals."

Reviewer#2, Concern # 10 Reference all figures including 4 and 5. Reference all tables.

Author response: We appreciate the reviewer's attention to detail.

Author action: In response to this feedback, we have revised our manuscript to ensure that all figures, including Figures 4 and 5, as well as all tables, are referenced appropriately within the text.

Reviewer#2, Concern # 11:

Not a complete sentence!

Define all variables including p and m.

Define all variable such as beta, epsilon, and gamma.

Change from all Caps

Author response: We would like to thank the reviewer for the constructive suggestion.

Author action: We modified the sentence :

“We demonstrate that deep neural networks can accurately classify the number of people within the Line of Sight (LoS), achieving an 99\% performance and efficiency with respect to both memory size and FLOPs (Floating Point Operations Per Second).”

“Here, p represents a specific position in both the input and output signals. The symbol w is used to denote the convolutional kernel, while and m is the size of the kernel.”

“Two common types of pooling are max pooling and average pooling which are introduced in ~\cite{maxpool, avgpool1, avgpool2} and presented in Table \ref{tab:pooling_equations}.”

“In this work, we experimented with both batch normalization and layer normalization, as introduced in \cite{ba2016layer, ioffe2015batch} and presented in Table \ref{tab:normalization_equations}.”

“In these equations, β represents the shift parameter, ϵ is a small constant added for numerical stability, and γ denotes the scale factor.”

“The Journal of the Korean Institute of Electromagnetic Engineering and Science”