On the use of pre-trained image classifiers for fingerprint-based indoor localization

NIEK PENNINGS, University of Twente, The Netherlands

Fingerprinting is a popular technique for indoor localization. It allows for the use of already existing infrastructure and offers satisfactory precision. The main drawback of fingerprinting is the arduous preparation time of gathering the fingerprint. Depending on the size of the building, this can be a big part of setting up an indoor localization system. In this paper, we will evaluate the precision of using pre-trained image classifiers for our training and testing. First, we will turn the fingerprint into an image, then it is ready to be used for training. Although the use of pre-trained networks can save time, their use and accuracy are not comparable to custom architectures.

Additional Key Words and Phrases: Indoor Localization, Pre-trained image classifiers, Transfer Learning, Fingerprinting

1 INTRODUCTION

Indoor localization is the process of determining a device indoors with satisfactory accuracy. Depending on the use case satisfaction can range from 1-2m to more. In the beginning, indoor localization was also done by using the already existing GPS system. The advantage of this was that for outdoor localization GPS has always worked quite well. Next to that, the infrastructure was already in place. Later on, new infrastructure will in some cases also prove to be a necessity for indoor localization. The disadvantage, and the main reason for improvement, is that GPS is unable to differentiate between floors. GPS can determine with some accuracy the position in a flat building, but as the building starts to rise, the accuracy will decrease. Therefore new techniques were needed for more precise indoor localization. As it turns out, various industries have expressed the need for better indoor localization. Whether it is in the health sector or in surveillance, being able to identify devices with an accuracy of several meters on different floors has become a necessity[3][16][10][11].

Nowadays different techniques like received signal strength indicator (RSSI), time of flight (ToF), angle of arrival (AoA) and technologies like WiFi, Bluetooth, ultra-wideband (UWB) and RFID are being used. Given these techniques and technologies, there are approximately five methods for localization, namely trilateration, triangulation, fingerprinting, Centriod and DV Hop. The latter two are fairly new and not yet fully developed. Fingerprinting is one of the more popular methods [13]. Lymberopoulos *et al.* [12] have also shown that fingerprinting is one of the more accurate techniques for indoor localization. Although they also concluded that methods like triangulation in combination with time of flight are also accurate. However, a choice has been made to focus on fingerprinting rather than trilateration or triangulation. Trilateration or triangulation require better calibrated or complex hardware [16] and focus more on data preparation and calculation. On the other hand, fingerprinting relies a lot on the fingerprint database it constructs and the data it uses for the prediction from this database, which means it is more focused on the quality and quantity of the data [13]. This paper seeks to use, among other techniques, transfer learning which is a technique that transfers knowledge from other source data sets into the model for the target data set. Hence, fingerprinting is the preferred method for this paper. [10]
For fingerprinting, either RSSI or channel state information (CSI) can be used. Although CSI gives a more accurate estimation of the received signal strength, it requires more complex hardware with several antennas. More and more antennae nowadays are able to use CSI, but not all of them. RSSI is very popular due to its low complexity and the no extra requirements. The trade-off however is that RSSI is very prone to reflections and inferences. This means that the data needs to be analyzed very carefully. Nonetheless, it will still be the method of choice in this paper. [16].

2 PROBLEM STATEMENT

There are not only upsides to using fingerprinting in combination with RSSI. Li *et al.* [8] have established the two main problems for this configuration. The first problem is *signal variation.* The received signal strength fluctuates a lot depending on various factors like humidity, human movement or automatic power adjustment strategies of the device. These can be generalized as signal reflection, non-line of sight and object movement [10]. High signal variation means that the RSSI measurements with the same device and on the same spot, can differentiate from day to day or even from hour to hour [8]. This can lead to problems with data collection and prediction. A subset problem from signal variance, that is significant enough to be mentioned separately is *database deterioration.* This means that although your model may have initially been quite accurate, in due time the building might change due to renovations which either cause disturbance in the RSSI or invalidates/outdates the current fingerprint [10].

The second problem is *device heterogeneity.* These indoor localization systems have a lot of use cases for smartphones, however the antenna and chipsets from the various smartphone manufacturers are quite different. This results in significant changes in the RSSI, which complicates the collection of the fingerprint and the localization[15].

There is also a third problem which is not necessarily described by Li *et al.*, but something that is an inherent problem. Whenever collecting data, especially in an environment where precision matters a lot, you will see human errors. Although we aim to collect data of reasonable quality, it will mean that there will always be little errors in the collection. Lastly, this is a bit more specific for

TScIT 38, February 3, 2022, Enschede, The Netherlands
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

fingerprinting and that is time. Collecting a fingerprint, especially of a big building can demand a lot of effort. In fingerprinting there is the trade-off between high overhead for the collection of the data and high accuracy, or fewer data collection but, potentially, also a lower accuracy.

Given this set of problems, the following research questions are formulated.

- **RQ1:** To what extent do benefits from a learning-based approach lift inherent limitations regarding environmental and device dynamics?
- **RQ2:** In view of the supervised learning manner that the existing fingerprinting techniques follow, it is implied that the more training data, the better the localizing performance. Which effective strategies mitigate the data burden but augment the localizing performance, or at least trade off the laborious costs well against the generality of the model?

3 RELATED WORKS

In some of the older research, Sun *et al.* [14] propose a transfer learning algorithm that tries to detect the underlying patterns by the data sets in a low-dimension space. Based on this, the target domain doesn't need to label all the data. However, it turns out that the performance is still impacted a lot by the amount and locations of the unlabeled data. In reference [17], the problem of not having enough target domain data is addressed. TrAdaBoost is a framework that only needs a little bit of target domain data in combination with old data or source data [1]. Zhang *et al.* [17] use the TrAdaBoost system in corporation with two techniques called One-Hot encoding and One-vs-Rest algorithm to prepare the data such that TrAdaBoost can use it. Li *et al.* [9] propose a framework called 'TransLoc' which filters and cleans the source domain data, then tries to find a cross-domain mapping between both the data sets. When that has been achieved and a homogeneous feature set is left, the mapping and transfer weights become a joint objective function for the training. This is useful for heterogeneous feature spaces.

In reference [4] a comparison is made between a custom CNN network and pre-trained networks. First, all of the fingerprints are transformed into an image by making a matrix and then normalizing the data. More than a dozen different models were evaluated with their own training data and prediction layer. Although their own custom CNN had arguably the best results, some other options gave similar results. It highlights the strength of CNN in this field, but also that transfer learning can be a good alternative if there is not a lot of data in the target domain. Li *et al.* [8] also transform the fingerprint into an image first and then run it through their custom CNN architecture. One of their focus points was to create a model that is robust and that shows little degradation over time. Their way of turning a 1D array fingerprint into an image is a technique that will later on also be used in this paper. A different way of turning a fingerprint into an image is shown in [7]. Laska *et al.* describe a way to also include the centrality and importance of the access points (AP) in the image. This is done by placing the closest and most important AP in the middle and then greedily surrounding the middle one with less important APs. Jang *et al.* also turns their fingerprint into a square and then provide a custom CNN architecture [6].

4 METHODOLOGIES

Nowadays, a few approaches exist for using CNN in classification problems. As described by [4], there are upsides and downsides to each of the different approaches. They mention the use of transfer learning for such a classification problem. In this case, an existing and pre-trained network is used to identify basic features like colors and shapes. The old classification layer is replaced by one suitable for the new classification problem. This network is then trained again. This way the knowledge of how to classify a picture is already there and this saves time that was otherwise needed by training the entire network from scratch. In this paper, this process will be the approach. By saving time by using readily available knowledge, one of the main drawbacks of fingerprints, namely the effort and time of creating the fingerprint and then training it, can potentially be reduced.

4.1 Preparation

As is a common problem with many machine learning models, the quality of the input data influences the quality of the results. Therefore the data needs to be processed correctly, however it is not quite apparent what is correct. The data of a fingerprint is in the form of a 1-D Array of signal strengths from the different access points. Because the approach of pre-trained CNN requires pictures, the initial fingerprint first needs to be transformed into an image. In section 3, some of these pre-processing techniques have been mentioned. Although there are many techniques out there, it's not necessarily clear which is the best.

In [8] a method has been proposed to do this in a certain way, such that the image is constructed with the relevant information. For the creation of the 2-D Matrix, the Nth row is dependent on the Nth element in the initial array. It is created as follows

$$\mathbf{x} = \left\{ \begin{array}{cccc} f^*_{1,1} & f^*_{1,2} & \cdots & f^*_{1,N} \\ f^*_{2,1} & f^*_{2,2} & \cdots & f^*_{2,N} \\ \cdots & \cdots & \cdots & \cdots \\ f^*_{N,1} & f^*_{N,2} & \cdots & f^*_{N,N} \end{array} \right\} \tag{1}$$

where

$$f^*_{j,k} = (f_j - f_k) * \frac{1}{f_k}$$

is the function that defines the elements [8]. One of the reasons that this method has been chosen for the creation of an image, is that it creates relevance for all the elements in their neighborhood. For image classification, CNNs infer knowledge by looking at the neighboring pixels. This means that the creation of the image can not arbitrarily be made by squeezing down the fingerprint into an image [4]. For example, if the fingerprint has N elements such that $\sqrt{N} \in \mathbb{Z}$, a square can be made from the 1-D array. However this naïve approach does not include the locality and important neighborhoods from the different APs. This naïve approach does include all the features of the original data. By combining the signal strength of the different measurements for the rows of the image, it can be that unnecessary noise is created.

During testing, it turns out that this approach as proposed by Li *et al.*, was the most successful. For convenience's sake, this approach has been called iToLoc. Two other approaches have been compared as well. The initial fingerprint consisted of 24 signal strengths. If one 0 is added it becomes 25 which is flattened to a square of 5x5. Another approach that has been compared is that this 5x5 is then tiled six times to create a 30x30 where this 5x5 is represented 36 times. After an image is created, the original values, ranging from -110 to -43 dB, are normalized to numbers between 0 and 1. Because these are pixels and represent colors, it is multiplied by 255. Afterward, the image needs to be resized to the dimensions required by the models, like 32x32, 224x224 or 299x299. These models also require a full RGB code, which means three dimensions. The resizing and the adding of extra dimensions are done by a CNN functioning as a data augmenter with only two layers, one for resizing and one for adding the extra dimensions. The upside of using the 5x5 is that all of the core features are represented in the image. The downside is that the CNN is going to upscale the image which can potentially create noise or unwanted features. The advantage of the 30x30 is that less upscaling is needed by the CNN, however the tiling can also create different unwanted features or patterns. It was important for the CNN what pixels are next to each other, however right now arbitrary pixels are placed next to each other.

4.2 Training

After the data preparation is done, the models for each of the different pre-trained models are built. When these models are used, the last few layers that are involved in the prediction are removed. These layers for most of the models were trained to classify images from ImageNet, however the fingerprint images are nothing like the image trained for ImageNet, so they are not usable. Instead, a custom prediction layer is added to each of the models. In the case of the resnet models and wifinet and alexnet that means the prediction layer was defined as follows.

- Global average pooling layer 2D
- drop-out layer (0.5)
- Dense layer (2048)
- Dense layer (2)

The head for the other models was slightly different, they did not contain a drop-out layer and the dense layer was 1024 instead of 2048.

- Global average pooling layer 2D
- Dense layer (1024)
- Dense layer (2)

When the training of the prediction has been done, the last two layers of the pre-trained model are unfrozen and fine-tuned such that they are adapted for the use of the fingerprint image.

5 RESULTS AND DISCUSSION

5.1 Experimental setup

For the training of the models new data was collected. Only one experiment has been performed to gather the data. It was not possible to give attention to *database deterioration* in this paper by gathering more data spread across different periods. For the collection of the data, two Samsung S8 phones and four HTC U11 phones have been used. The Samsung phones have only been partially used due to their performance hindering the collection. For the sake of having more data, their data has been kept. The rooms and data collection points can be found in 5 in the Appendix. On the floor in this office, three APs have been hung from the ceiling, they provide the required connectivity for the collection. Each of these APs has eight directional antennas. This allows for great variation in the signal strength received from all of the APs. The only thing these APs do is receive a Bluetooth signal and then bounce it back. The phone then gathers this signal and saves it. The schematic of one of the APs can be seen on 1. Because there are three APs and each AP has eight antenna's, the fingerprint in the data consists of 24 signal strength measurements. For all of the points on the map that have been measured, the fingerprint was collected in every wind direction. This ensures that it does not matter in which direction the human collecting the data was standing. The human body would absorb a part of the Bluetooth signal propagated by the phone [5].

Two different metrics are presented in 3 to evaluate the information they both present. As can be confirmed in 2, iToLoc gave on average the best performance for most of the models, therefore it is also used in the comparison between the MSE and MAE.

Table 1. Comparison in MSE and MAE

Model	MSE (m)	MAE (m)
vgg19	9.2369	2.8039
inceptionv3	14.1996	4.4854
convnextbase	14.7303	3.8743
resnet50_v2	18.1500	3.7698
wifinet	18.4849	2.4909
resnet152_v2	19.7764	3.9597
resnet101_v2	27.0536	3.8857
efficientnet_v2	29.8050	4.9081
alexnet	43.7641	4.1204
resnet50	44.7586	4.6963
resnet152	59.8317	4.6918
resnet101	76.7442	4.5262

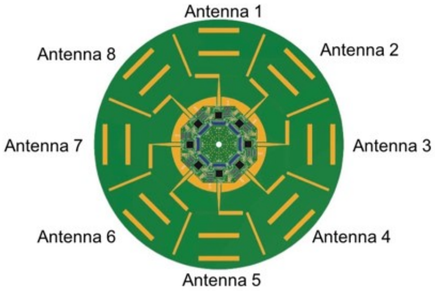

Fig. 1. The schematic of one access point [2]

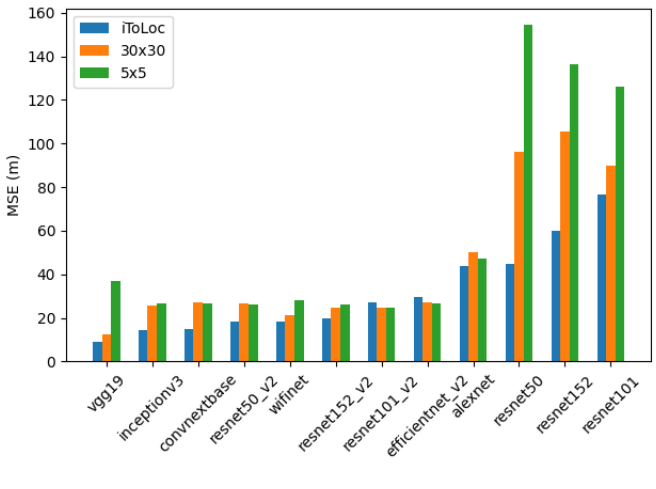

Fig. 2. Comparison in image creation techniques

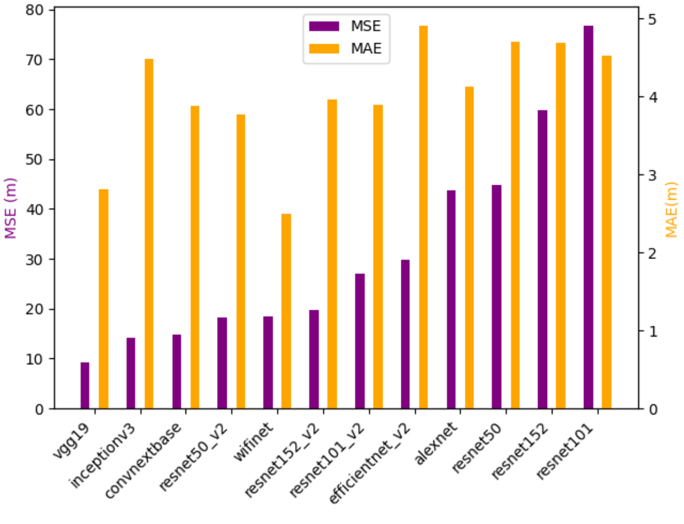

Fig. 3. Comparison in MSE and MAE

5.2 Results

The results can be found in figure 2, figure 3 and table 1. As can be seen that the performance varies a lot, with the worst performance at 76.7442 m and the best performance at 9.2369 m for the MeanSquaredError. The best performance and the worst performance when measuring with MeanAbsoluteError range between 2.8039 m and 4.9081 m. MAE shows us some of the potentials for a practical application. It tells that low errors are possible. However, the highs in the MSE show that there are still many errors. Because of the square in the metric, outliers as seen in the figure and table are highlighted. This indicates that the performance still needs to be optimized.

Although about 9 m for MSE could potentially be acceptable for room detection. It could provide the room or the room next to it to the user. In the case of a shopping mall or another very generic use case, it would be acceptable. However in the case of room classification where high accuracy is needed, this would already be insufficient. The best-performing model gives unsatisfactory results if there is a requirement of a few meters. By comparison, iToLoc [8] gained a mean accuracy of 1.86m and a 95th percentile of 5.41m, WiFinet [4] gained a Root Mean Squared Error (RMSE) of 28 cm and ResNet18 with Transfer Learning gained an RSME of 24.6 cm.

The saliency map in 4 shows us what is important to the model. It turns out that for this instance of the picture, only the inner parts of the picture matter. The data in the middle square is mostly related to the information from the middle access point. This means that the data from the middle access point is the most important. It is good to note however that there is perhaps a bias in these models for detecting the relevant information in the middle rather than the entire picture.

To be able to critically analyze whether learning-based approaches are able to combat environmental and device dynamics, more research needs to be done. Right now it is only known what the results are when almost all of the data is from one device and from one measurement. The same approach needs to be taken with more dynamic data if a structured answer is to be given to RQ1.

Although, depending on the requirements, it is inconclusive if the usage of pre-trained models is a good approach for indoor localization. Many models show MSEs of above 20m which is a lot and in most cases would probably be far from ideal. As it stands right now, it seems that it cannot be a valid strategy. It could however also be the quality and/or quantity of the data. Perhaps the lower bound or the minimum of data required for accurate enough results is higher than what has been collected in this paper. Only two days have been spent collecting data and many optimizations can probably be made (see Future works).

6 FUTURE WORKS

Due to the nature and length of this paper and the generality of the research, there are a few improvements for future research. These are highlighted and touched upon in this section. The first improvement is in pre-processing. There are not that many different techniques out there for creating an image from a fingerprint and many techniques are naïve or do not elaborate much on the process. Because the fingerprint and the data are so defining, this could potentially be an area that sees more research. The saliency map can provide more insights into how such an image should be made from the fingerprint. Next to that, the data from the saliency map in 4 shows that by cutting off the edges, the data is already fine-tuned a bit more towards what is important. This gives the CNN less noise and more accurate information.

It is not always possible to test and try every combination. More models can be tried and the exact configuration of running these models can be improved. All of the models are run on the same configuration, however this means that the prediction layer(s) are more suitable for one model than another. More effort can be put into finding models that are more suitable for this kind of work.

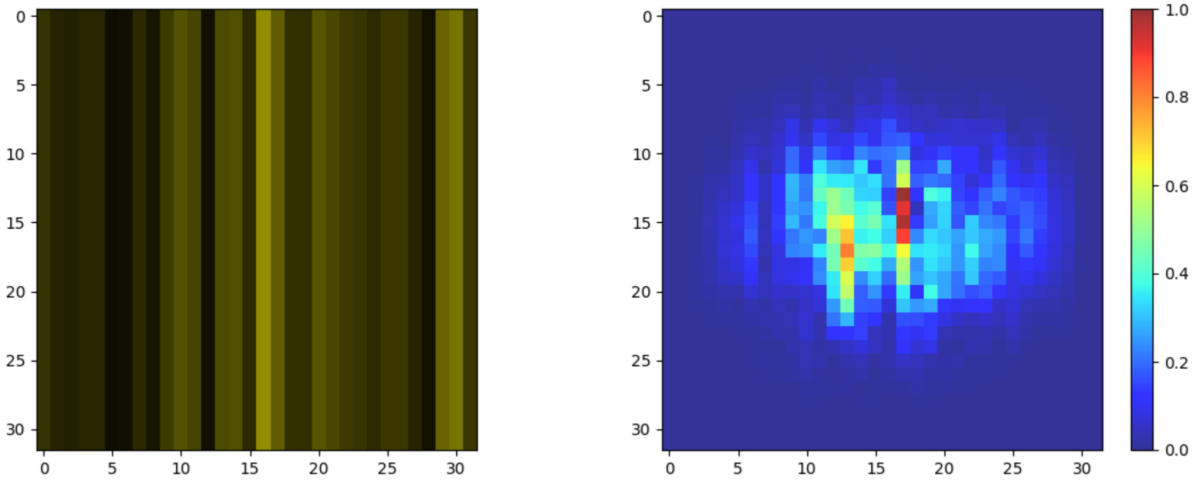

Fig. 4. The visualization and the saliency map of a picture

The random and human factors also still play a role in key parts of this research. The data set might contain errors that can lead to worse results. Some of the dimensions of the data are also generated by an AI, which means that there is a random factor in that. Sometimes the data might be generated in such a way that some models might perform better than others times. This can also happen the other way around. To avoid this human error, other publicly available datasets can be used in the comparison.

Lastly, given the fact that just using pre-trained models falls short of other approaches, it could be applied to purely room classification. For many applications, room rather than exact position classification can already provide support. When the best performance of about 10 m is given, room classification might be able to reach high accuracy. The way the data was collected in this paper, it was not possible to study or evaluate that.

REFERENCES

[1] Dai, W., Yang, Q., Xue, G. R., and Yu, Y. Boosting for transfer learning. In *ACM International Conference Proceeding Series* (2007), vol. 227.

[2] de Haan, T. BLE Localization Using Switched-beam Angle Of Arrival for Pallet Localization In Warehouses.

[3] Hameed, A., and Ahmed, H. A. Survey on indoor positioning applications based on different technologies. In *12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, MACS 2018 - Proceedings* (2019).

[4] Hernández, N., Parra, I., Corrales, H., Izquierdo, R., Ballardini, A. L., Salinas, C., and García, I. WiFiNet: WiFi-based indoor localisation using CNNs. *Expert Systems with Applications 177* (2021).

[5] Hong-xin, Z. Specific absorption rate distribution in human body caused by PIFA for bluetooth applications. *Chinese Journal of Radio Science* (2008).

[6] Jang, J.-W., and Hong, S.-N. Indoor Localization with WiFi Fingerprinting Using Convolutional Neural Network. In *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)* (7 2018), IEEE, pp. 753–758.

[7] Laska, M., and Blankenbach, J. Topology Preserving Input Image for Convolutional Neural Network Based Indoor Localization. In *2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (11 2021), IEEE, pp. 1–8.

[8] Li, D., Xu, J., Yang, Z., Lu, Y., Zhang, Q., and Zhang, X. Train once, locate anytime for anyone: Adversarial learning based wireless localization. In *Proceedings - IEEE INFOCOM* (2021), vol. 2021-May.

[9] Li, L., Guo, X., Zhao, M., Li, H., and Ansari, N. TransLoc: A Heterogeneous Knowledge Transfer Framework for Fingerprint-Based Indoor Localization. *IEEE Transactions on Wireless Communications 20*, 6 (2021).

[10] Liu, K., Zhang, H., Ng, J. K. Y., Xia, Y., Feng, L., Lee, V. C., and Son, S. H. Toward Low-Overhead Fingerprint-Based Indoor Localization via Transfer Learning: Design, Implementation, and Evaluation. *IEEE Transactions on Industrial Informatics 14*, 3 (3 2018), 898–908.

[11] Long, K., Zheng, C., Zhang, K., Tian, C., and Shen, C. The Adaptive Fingerprint Localization in Dynamic Environment. *IEEE Sensors Journal 22*, 13 (2022), 13562–13580.

[12] Lymberopoulos, D., Liu, J., Yang, X., Choudhury, R., Handziski, V., Sen, S., Lemic, F., Busch, J., Jiang, Z., Zou, H., Pirkl, G., and Hevesi, P. A realistic evaluation and comparison of indoor location technologies: Experiences and lessons learned. In *IPSN 2015 - Proceedings of the 14th International Symposium on Information Processing in Sensor Networks (Part of CPS Week)* (2015), pp. 178–189.

[13] Nessa, A., Adhikari, B., Hussain, F., and Fernando, X. N. A Survey of Machine Learning for Indoor Positioning. *IEEE Access 8* (2020).

[14] Sun, Z., Chen, Y., Qi, J., and Liu, J. Adaptive localization through transfer learning in indoor Wi-Fi environment. In *Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008* (2008).

[15] Tiku, S., Gufran, D., and Pasricha, S. Multi-Head Attention Neural Network for Smartphone Invariant Indoor Localization. In *12th International Conference on Indoor Positioning and Indoor Navigation, IPIN 2022* (2022).

[16] Zafari, F., Gkelias, A., and Leung, K. K. A Survey of Indoor Localization Systems and Technologies. *IEEE Communications Surveys and Tutorials 21*, 3 (2019), 2568–2599.

[17] Zhang, Y., Wu, C., and Chen, Y. A Low-Overhead Indoor Positioning System Using CSI Fingerprint Based on Transfer Learning. *IEEE Sensors Journal 21*, 16 (2021).

A APPENDIX

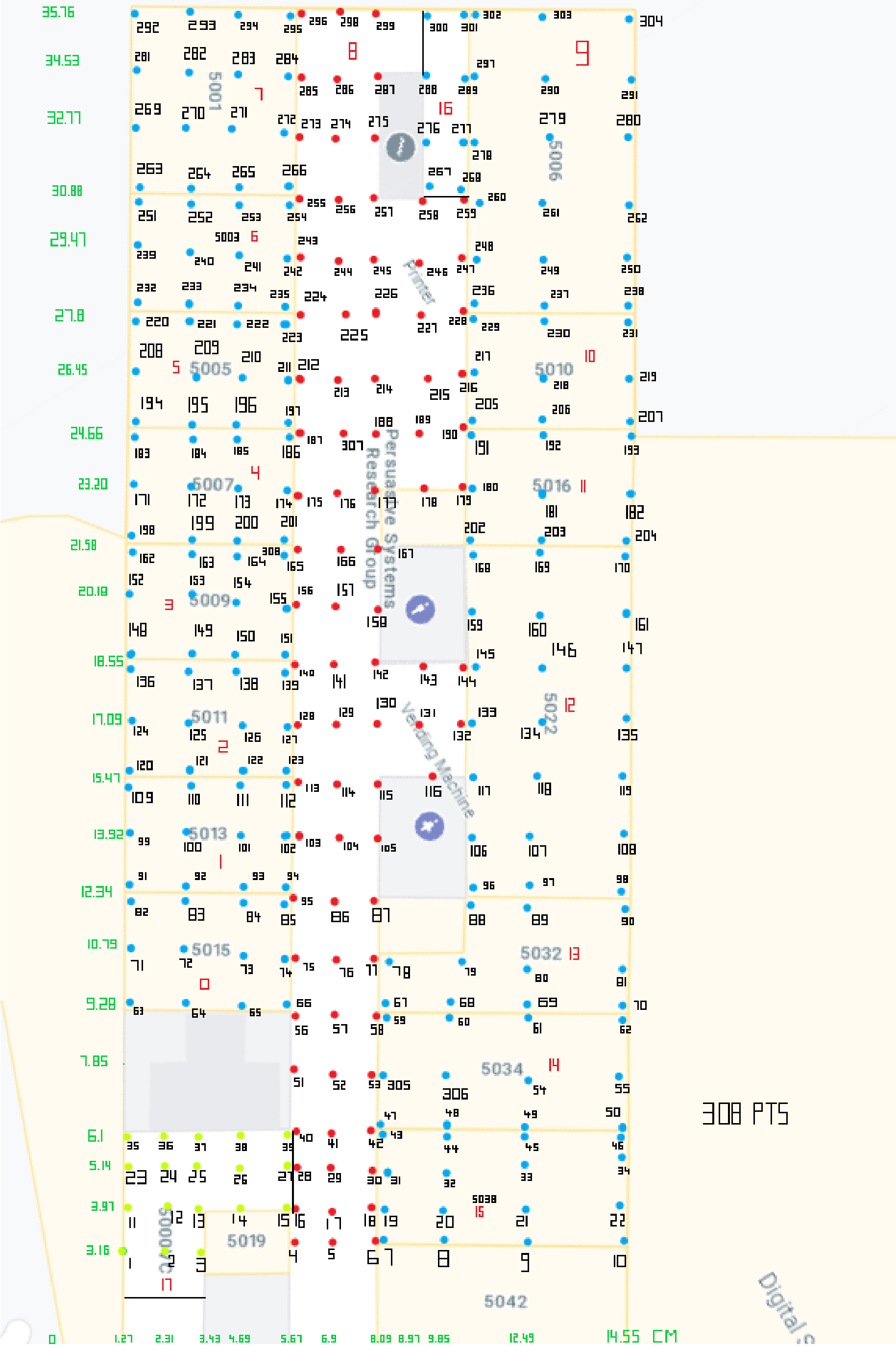

Fig. 5. The map of the office building. Each of the points, regardless of the fact of they are blue, red or green, shows a point collected in the data set. The numbers on the 'x-' and 'y-axis' show the coordinates of each of the separate points. The red numbers in the rooms give a room number to each of the coordinates. The access points are not shown on this map.