Toward Multi-area Contactless Museum Visitor Counting with Commodity WiFi

YICHENG JIANG and XIA ZHENG, Zhejiang University, School of Art and Archaeology, China
CHAO FENG, Northwest University, School of Information Science and Technology, China

Multi-area visitor counting plays a critical role in museum management, which can help administrative staff better study visitor flows and hotspots, so that they can ensure the quality and safety of visits. Internet of Things (IoT) techniques facilitate efficient recording and understanding of visitors' spatial and temporal distribution in museums, and traditional visitor tracking applications use surveillance cameras or wireless connections with portable smart devices. However, these methods either involve privacy concerns or face the obstacle of getting natural behavioral data of all visitors. This article explores an IoT monitoring methodology in the field of museum studies, proposing a commodity WiFi-based head-counting framework that does not need the visitor to connect with any device. Our system analyzes the Channel State Information amplitude fluctuations at the fixed receiver caused when visitors cross the line-of-sight link. It enables multi-area visitor counting by achieving In-and-Out traffic detection at different sites with a convolutional neural network algorithm. The method also allows for a rough classification of visitor types based on body size, and an extra transfer module is presented to reduce training time for increasing sensing scenarios. Over 2,300 samples at five different sites were collected to test the usability. Experiment 1 implemented in three environments/deployments demonstrated that the proposed approach can be potentially implemented in variable sites of museums. It achieved high up to 95% and 99% accuracies for identifying the number and direction of people crossing, respectively. Experiment 2 sampled adults, children, and adult-child groups at a science museum and achieved approximately 89% classification accuracy of visitor types. Experiment 3 collected data for all cases in which up to three targets entered and exited simultaneously, and reached a recognition accuracy of around 88% for nine different cases. The potential and limitations for the practical application of wireless contactless sensing to cultural spaces are discussed.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing;**

Additional Key Words and Phrases: Commodity WiFi, neural network, flow counting, visitor study

ACM Reference format:
Yicheng Jiang, Xia Zheng, and Chao Feng. 2023. Toward Multi-area Contactless Museum Visitor Counting with Commodity WiFi. *J. Comput. Cult. Herit.* 16, 1, Article 8 (March 2023), 26 pages.
https://doi.org/10.1145/3530694

1 INTRODUCTION

Understanding the visitor traffic is central to museum operation, it can be of benefit to both visitors and museum administrators. On the one hand, knowing the real-time crowd situation and historical hotspots of the

This research was supported by the National Key Research and Development Program of China (Project No. 2019YFC1521105).
Authors' addresses: Y. Jiang and X. Zheng (corresponding author), Zhejiang University, School of Art and Archaeology, Hangzhou, China; emails: {jiangyicheng, zhengxia}@zju.edu.cn; C. Feng, Northwest University, School of Information Science and Technology, Xi'an, China; email: chaofeng@stumail.nwu.edu.cn.

1556-4673/2023/03-ART8 $15.00
https://doi.org/10.1145/3530694

museum helps visitors learn about popular exhibits and adjust their visit plans, improving the visitor experience. On the other hand, the administrator requires to know the visitors' points of interest, to better provide services like exhibit recommendation and route design, as well as accumulate experience in follow-up curations.

Visitor traffic detection in museums is driven by a simple question: how to obtain high-quality, large-scale data with the informed consent of the privacy-sensitive visitors? Many efforts related to **Internet of Things (IoT)** have been made to answer the question, which can be mainly divided into two categories, device-based (contact) and device-free (contactless), depending on whether the visitor needs to carry the device or not. Some portable smart devices, such as smartphones, GoPros, or guides carried by visitors [21, 50, 56] could calculate the location or nearby exhibits in real time and transmit the information by wireless signal, but not all visitors are willing to accept the device-based solutions due to the burden of carrying or potential privacy risks. In contrast, device-free methods attract more attention, since they do not require the target to wear any device. For example, surveillance cameras can be used to efficiently extract the number or route of the targets [25, 62, 87], which however is sensitive to obstacles and angle conditions, hindering the sensing accuracy. Although some other methods, such as the **Passive Infrared (PIR)** sensor [75] and **Ultra-wideband (UWB)** [11], are less affected by light and occlusion and can achieve accurate and privacy-preserving counting, large-scale deployments in various environments may lead to relatively high costs.

To meet the requirements of desirable accuracy and scalable deployment of visitor traffic sensing, this article attempts to design and test an IoT smart space technique that has not been fully experimented with in the museum environment. Specifically, we propose to leverage **commercial-off-the-shelf (COTS)** WiFi devices to achieve contactless visitor counting in multiple sites of a museum. COTS WiFi transceivers are very cheap and can be widely deployed in museums. Fine-grained **Channel State Information (CSI)** is easily accessible on commodity WiFi devices. Recently, many researchers have proved that WiFi CSI can be used to sense target activity based on the fact that different activities cause different signal changes [32, 55, 77, 78, 89]. Although WiFi-based device-free sensing technology offers a natural, privacy-preserving, and low-cost possible solution, framework design and field tests in museums are relatively scarce. Therefore, this article aims to implement an integral scheme for multi-area visitor counting, making up the gap between the general WiFi-based contactless sensing approach and the multi-area museum visitor counting. Our high-level idea is to construct the unique relationship between visitor traffic and CSI amplitude change.

However, translating our high-level idea into a practical system requires us to overcome several significant challenges. First, most areas in a museum are opened and interconnected, people who come to one site may moves to another site at the next moment. So it is difficult to decide where to deploy the WiFi transceivers. The past attempts choose to place the transceiver in a certain area to detect the crowd size inside [14, 43, 80, 92], which while lacking clarity in monitoring spatial boundaries, thus hinders the application of open multi-region sensing. For instance, if more than two pairs of transmitter and receiver devices are deployed, respectively, in adjacent rooms, then the current counting models may be disturbed by the visitors in neighboring areas. Therefore, given the particularity of the museum environment, this article proposes to achieve visitor counting by flow counting at the door between any two adjacent sites. Specifically, the problem of multi-area visitor counting could be converted into judging the number of people passing the door between neighboring regions. Considering the change in WiFi signals caused by persons crossing the **line-of-sight (LoS)** is greater than those caused by other walking directions, the transmitter and receiver are separately deployed in pairs on the connection location between every two zones (see Figure 1), which allows the location of the devices to be the boundary of artificially defined areas.

The next question is how do we achieve accurate flow counting? To answer this question, existing systems employ a similar deployment for counting the passing people [18, 91], while not yet having implemented direction judgment. If we only count the visitors without identifying whether they are entering or leaving, then there

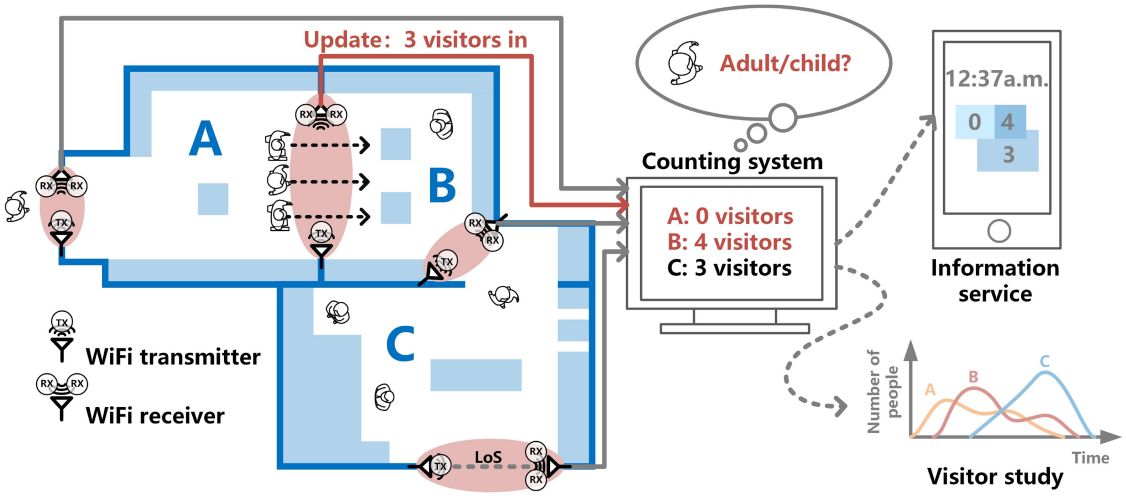


Fig. 1. Example application of the proposed system.

will be a severe misjudgment in identifying the head count inside the area. In this article, we aim to design an algorithm that is capable of recognizing both the number and direction of people passing like [17, 83, 84] in a real museum scenario. We associated two receiving antenna amplitudes over time with the ground truth using a two-channel **convolutional neural network (CNN)**, to extract unique features for head count and direction recognition. Furthermore, the network is expected to identify the type of the visitors passing through, such as an adult or a child, given the disparity in the signal interference between people with large and small body sizes. For cultural space scenes, the detection of simultaneous moving in and out is also practical.

Finally, we assume the subsequent renewal of the system. The flexible layout of multiple museum spaces determines the diversity of equipment deployments. When the system is required to improve the ability to identify more head count or adapt to a wider range of situations, for each deployment point all parameters need to be totally retrained with new samples, which can take a lot of time. As a result, the reusability of feature extraction parameters is taken into account. This article additionally introduces pre-training and fine-tuning to minimize the training cost when the identifiable cases increase.

Contributions. To summarize, our main contributions can be summarized as follows:

- This article designs a framework for contactless multi-area visitor counting in museums with commodity WiFi. Each monitoring site requires only a cheap router and a mini-pc with two antennas. To the best of our knowledge, it is the first study that implements human-centered device-free WiFi sensing in museums.
- This article proposes a two-channel CNN model to identify the number, direction, and visitor type of passing people by extracting unique features from CSI measurements. A transfer learning approach is involved to reduce the training time for the addition of new samples.
- We evaluate the system performance in multiple real museum scenarios. Intensive experimental results show that the system can achieve up to 95% and 99% accuracies for head count and moving direction recognition, respectively. The pre-training and fine-tuning of the CNN model structure could reduce the average training time by about 75% while ensuring that the average accuracy difference is stable within 6.25%. The method's accuracy for classifying adult, child, and adult-child group was nearly 89%, while its accuracy for identifying simultaneous entry and exit circumstances involving up to three individuals was about 88%, demonstrating the framework's potential and limitations.

Table 1. Summary of Timing and Tracking Technologies in Museums

		Space coverage cost	Visitor coverage cost	Privacy protection cost	Accuracy	Applied or experimented in museums
Vision-based	Device-based	Low	High	High	High	[19, 53, 56, 60, 69, 70]
	Device-free	High	Low	High	High	[25, 27, 28, 87]
Non-vision-based	Device-based	High	Moderate	Low	High	[10, 21, 30, 33, 38, 49, 50, 59, 66, 67, 74, 85]
	Device-free	High	Low	Low	Low	Very few

2 RELATED WORK

2.1 Timing and Tracking Technologies in Museums

Digital technology gives museums new ways to engage with visitors, and the growth of IoT applications in smart environments makes embodied and tangible interaction between visitors and objects in museum space a main focus [47]. One of the most important technological needs for visitor management and services is understanding the time and space distribution behavior of visitor flows [41]. Lanir et al. interviewed a series of museum stakeholders and derived five needs for analysis of visitor behavior: visitors' engagement at the room and exhibit level, visitors' flow and pattern between rooms, time analysis of individuals and small group visits, demographic and time-based segmentations, and organized groups and educational activities [38]. In general, museum visitor behavior analysis requires collecting all-time records and spatial logs at optional granularities (room/exhibit-level, group/individual-level, etc.).

Visit time is one of the most critical measurements in visitor study. In 1997, Serrell proposed that the indicator of the visitor's learning could be measured equally to the visit time and stops in front of different exhibits, helping curatorial teams to better decide the exhibition size and media format [61]. In the study of Emerson et al., dwell time was used as a measure of engagement and as a target for multimodal prediction tasks [20]. Time spent on exhibits can be related to many variables. Johnson's regression analysis involving over 50 variables on the visit time of 501 visitors at six zoos found that the physical features of the exhibits and spaces were the main and significant influencing factors [37]. Other studies also explored associations between visit time and visitor group size, visitor demographics, museum fatigue, space and content design [4, 35, 64, 65, 71], and so on, which provide empirical data-based recommendations for museum curation.

Time analysis is inseparable from the observation of visiting routes or locations, and IoT technology can replace manpower to achieve various scales of indoor sensing. The most traditional timing and tracking way is paper-and-pencil. Empirical study has shown that reactivity effects are negligible when there are no further interactions between subjects and observers after the cuing [12]. Besides, the individual-based observations could facilitate further interviews to obtain more information for an explanation. However, it may pose concerns about accuracy, labor cost, and influence on the visiting experience. Toward this end, the self-mapping of tourists may serve as a proper method for leisure environment evaluation [52]. Additionally, time and tracking technologies, such as video recording and wearable devices, are increasingly being employed for visitor observation [82], allowing for more diverse study methodologies with larger samples. These automated methods are technically part of the indoor human detecting, **indoor positioning systems (IPS)**, and **electronic travel aid (ETA)** services, among others. We provide a brief overview of the various possible timing and tracking technologies and specify the categories to which the proposed method belongs.

We divide these technologies into four main categories based on whether they are vision-based and whether the viewers must carry devices (device-based/contact versus device-free/contactless), as shown in Table 1, which reflect differences in museum space coverage cost, visitor coverage cost, privacy protection cost, and accuracy. Overall, when compared to visual methods, non-visual methods generally compromise accuracy for visitor

privacy security, while contactless methods sacrifice space deployment cost for user coverage cost compared to contact methods.

Vision-based device-based methods. It refers to the acquisition of egocentric views via a visitor's cell phone, wearable camera, or eye-tracking glass for individual-level positioning based on environmental images. Combining with the study of emotion mapping, a smartphone that automatically took pictures was employed to calculate the locations where the user's mood changed throughout the tour route [60]. Eye-tracking devices have also been used in field museum studies, providing finer-grained attention distribution data beyond localization [19, 70]. User data can be utilized for more than just analysis; it can also be used to provide services. Some image datasets of visitor perspectives in museums have been used in automatic object recognition [53]. According to vision-based recognition, Ragusa et al. proposed a framework for locating visitors in cultural sites by tagging egocentric images from wearable Hololens and GoPros [56]. Styliaras et al. implemented the MuseLearn Platform in Herakleidon Museum, where the system provided content recommendations for visitors viewing the exhibition via mobile devices [69]. These solutions simply entail the pre-entry of exhibit and environmental visual data without any changes to the cultural spaces. The approaches can surely produce high-quality data on individual user activity, but not all visitors are ready to use guide tour gadgets, which opens the door to contactless alternatives.

Vision-based device-free methods. In terms of device-free tracking in museums, surveillance cameras were first explored. Brunelli et al. built a virtual museum simulation environment to demonstrate the possibility of visual-based visitor tracking in large indoor environments [6]. Zabulis et al. designed a system based on camera networks in an archaeological museum to enable multi-person tracking and interaction in front of a large-scale display [87]. A single-camera multi-people tracking algorithm was proposed by Godbehere et al. and worked well in museum hours regardless of various lighting conditions [25]. In addition to typical RGB cameras, Infrared Radiation cameras can be employed in museum exhibits to recognize human behavior [28]. It is worth mentioning that time of stay per capita and average concurrent users are two types of metrics used for assessing quality and attractiveness. Therefore, a multi-regional head-counting system potentially provides a group-based perspective different from the individual-focused observation. However, the camera-based methods are sensitive to angle conditions and the obstruction by the exhibits, so they are mainly adopted in interactive installations [27] rather than in large-scale audience research. Besides, ethical concerns make it difficult to disclose the statistical results for public information services.

Non-vision-based device-based methods. It refers to connecting to a wireless signal for area positioning utilizing a guide or other smart device carried by the audience. Since many visitors already have smart devices that integrate wireless technology, device-based solutions are now more common. Moussouri and Roussos conducted a case study of family visitors in the London Zoo, illustrating the feasibility of using smartphones in tracking studies [50]. However, indoor environments such as museums demand greater precision in visitor tracking. A number of **Global Positioning System (GPS)**-based indoor positioning approaches have been proposed [40]. Besides, Sakpere et al. provided a systematic overview of current indoor positioning techniques [58], which include signal properties and positioning algorithms. The **Angle of Arrival (AOA)**, **Time of Arrival (TOA)**, **Time Difference of Arrival (TDOA)**, and **Received Signal Strength Indication (RSSI)** are examples of the former, while Triangulation, Trilateration, Proximity, and Fingerprinting are examples of the latter. Mahida et al. followed the flow of localization to sort out these parameters and algorithms in detail [46]. Brena et al. investigated the types of indoor positioning technologies [5], which consist of **Radio Frequency Signals (RF)**, light, sound, and Magnetic Fields. WiFi, **Radio Frequency Identification (RFID)**, BlueTooth, UWB, and so on, are all RF-based technologies. Specifically for museum applications, Escuer et al. evaluated a variety of RF-based tracking methods currently in museums [21]. Verbree et al. tested two methods of locating visitors based on WiFi devices at the Hubei Museum while discussing user concerns and legal issues regarding privacy protection [74]. Bluetooth data was instrumental in analyzing visiting patterns in the Louvre Museum [85]. While this strategy involves the deployment of a certain number of wireless devices in advance in the space, it can balance privacy

protection with service delivery. Signaling devices deployed in museum spaces can sense the location of visitors via smart guides [33, 59] or smartphones [10, 30, 49, 66, 67], providing interactive content services and enhancing the cultural experience of visitors. However, visitors who rent an electronic guide or connect a private device to the network still make up a small portion of the overall audience.

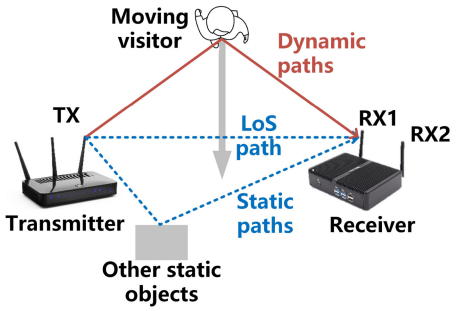
Non-vision-based device-free methods. RF-based wireless communication devices generally consist of transmitters and receivers (or tags and anchors). The signal types employed in the device-free approaches and the device-based ways are the same; however, the latter chooses to fix one of the two communication devices and use the other as portable equipment, whereas the contactless method fixes both in space and identifies possible situations based on changes in the signal caused by human activity. Therefore, contactless wireless sensing meets the demands for massive and privacy-insensitive information sharing with minimal disruption to the visitors. This obviously causes instability in recognition accuracy but can be compensated for by additional equipment, robust feature extraction algorithms, machine learning, and so on. Among the available options for RF-based contactless indoor tracking methods, WiFi is a more suitable medium considering universality, communication distance, and costs. He et al. proposed a comprehensive review of WiFi-based contactless sensing, indicating the prospects for its application in smart museums [32]. However, there are few frameworks designed and field experiments in real museums that address the needs of visitor studies.

2.2 WiFi-based Contactless Flow Counting

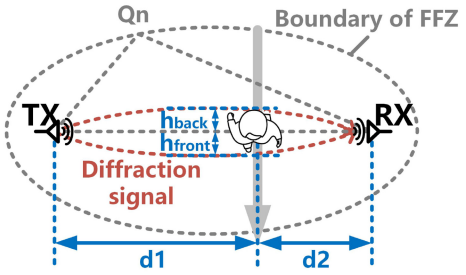
In the early stage of wireless sensing research, RSSI was mainly used as a feature. Lin et al. firstly exploited radio signal fluctuations for indoor automated people counting and explored the ability of single transmitter-multiple receiver at a distance of 1.5 m to detect two people [42]. Doong deployed a transmitter and two receivers with 2m spacing, identifying the number of people passing by under the condition of two opposite directions [17]. Depatla et al. deployed two pairs of transmitters and receivers to estimate the speed and number of a crowd when arriving/departing the area [13]. Compared to RSSI, CSI data has a finer granularity. Wu et al. leveraged WiFi CSI to detect a human's walking direction. The system related the direction of the dynamic target in the 2D Fresnel zone model to the phase change and achieved single-person direction monitoring with a median error of fewer than 10 degrees through a transmitter and two receivers [79]. These studies provide insight into human flow counting and direction estimation of the wireless signal. Nevertheless, the receivers in these studies were deployed in dispersed locations, which may not be suitable for spatially complex museum environments.

The development of artificial intelligence helps to further extract the deep features of the signal, resulting in the diversity of perceptible objects and the simplification of device deployment. Combining CSI data and **Deep Neural Network (DNN)**, Doong further monitored the number of up to five passers-by using a transmitter with one antenna and a receiver with three antennas [18]. Xiao et al. estimated queue size with WiFi via **Support Vector Machines (SVM)**, identifying queues of one to four people through a similar device with about 90% accuracy [81]. Zhou et al. innovatively leveraged the **Doppler Frequency Shift (DFS)** feature of WiFi signals to identify more complex cases in queues, i.e., a continuous flow composed of different subflows [91]. These efforts integrate wireless sensing and machine learning, broadening the horizon of WiFi-based flow counting. However, they do not imply both flow size and direction recognition for the time being, and our system reduced one receiving antenna based on this deployment.

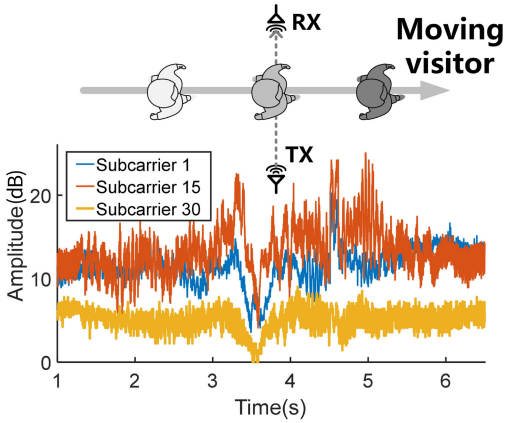
In terms of function, the closest to our study is Yang's work [83, 84]. The system used a transmitter and 2 adjacent receiver antennas, employing WiFi phase difference and the CNN model to achieve bi-directional head counting. The deployment placed the LoS perpendicular to the door, taking up a certain area in the room to some extent, so it may not be appropriate in public areas where exhibits need to be designed for placement. Our deployment still brings people through the LoS, because it is more in line with the museum, while also considering the case in almost completely different sites and system upgrades. We further experimented with the system's potential to identify the visitor type, and the simultaneous entry and exit of the targets.



(a) The reflection signals outside the FFZ.



(b) The diffraction signals inside the FFZ.



(c) The amplitudes of the 1st, 15th, and 30th subcarriers when a target passing.

Fig. 2. The RX signal consists of multiple paths.

3 PRELIMINARY

This section first introduces the format of the collected WiFi data, argues why dynamic targets cause changes in the signal flow, and then models the behavior of people crossing the link based on the Fresnel zone model.

3.1 Channel State Information

Considering both transmitter and receiver usually have more than one antenna, the antenna of the transmitter is abbreviated as TX and that of the receiver is abbreviated as RX. Figure 2 illustrates a TX-RX pair serving as the basic WiFi-based IoT device. The support of **orthogonal frequency-division multiplexing (OFDM)** and **multiple-input multiple-output (MIMO)** enables WiFi data to be transmitted as multiple subcarriers between different antennas. TX sends packets to RX continuously, and a CSI matrix can be extracted by the Linux 802.11n CSI Tool [31]:

$$H_{(i,j)} = \begin{bmatrix} h_{(1,1)} & h_{(1,2)} & \cdots & h_{(1,S)} \\ h_{(2,1)} & h_{(2,2)} & \cdots & h_{(1,S)} \\ \cdots & \cdots & \cdots & \cdots \\ h_{(T,1)} & h_{(T,2)} & \cdots & h_{(T,S)} \end{bmatrix}, i = 1, 2, \ldots, N_{tx}, j = 1, 2, \ldots, N_{rx}, \quad (1)$$

where N_{tx} and N_{rx} represent the total number of TX and RX antennas, respectively, $H_{(i,j)}$ is the CSI stream from the ith TX antenna to the jth RX antenna, consisting of S subcarriers with a length of T packets, and$h$ can be expressed as

$$h = |h|e^{j\angle h}, \tag{2}$$

where $|h|$ and $\angle h$ represent the amplitude and the phase, respectively.

As Figure 2(a) shows, the collected CSI stream is a mixture of signals from many different paths, which can be summarized into three categories: the direct signals on the LoS path, the reflection and diffraction signals caused by the dynamic people, and those caused by other static objects [8], e.g., walls, exhibits, and so on. When visitors cross the LoS vertically, the covered direct and static signals as well as the changed dynamic ones cause a fine-grained fluctuation in the amplitude and phase of all subcarriers.

3.2 Fresnel Diffraction Model

The Fresnel zone is an ellipsoidal region directly surrounding the LoS path. Wang et al. first introduced the model to the indoor environment for respiration detection via commodity WiFi [76], and Xiao et al. enhanced the stability of the human flow detection system by using the changes caused by people entering/leaving the Fresnel zone from different directions [81]. Considering the influence of the Fresnel diffraction, References [51, 89] studied the amplitude changes induced by human activity through the LoS. The Fresnel zones can be considered as a set of nested ellipses in the top view, the innermost of which is the **First Fresnel Zone (FFZ)**. The boundary of the nth Fresnel zone can be expressed as [8]

$$|TX, Q_n| + |Q_n, RX| - |TX, RX| = \frac{n\lambda}{2}, \tag{3}$$

where Q_n is the point on the boundary of the nth Fresnel zone, and λ represents the wavelength of WiFi.

The signal is dominated by reflection phenomena when the moving target is outside the FFZ and by diffraction phenomena when it is inside the FFZ as Figure 2(b) shows. The signals reach RX from both sides of the object, contributing to the gain:

$$S_{dif} = A(v_{front})e^{-2j\phi_{dif,front}} + A(v_{back})e^{-2j\phi_{dif,back}}, \tag{4}$$

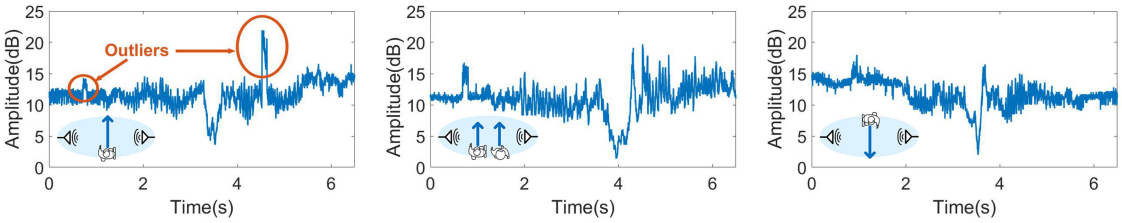
$$A(v) = \frac{1+j}{2}\int_{v}^{\infty} e^{\frac{-j\pi x^2}{2}}dx, \tag{5}$$

where $v_{front} = h_{front}\sqrt{\frac{2(d_1+d_2)}{d_1 d_2}}$ and $v_{back} = h_{back}\sqrt{\frac{2(d_1+d_2)}{d_1 d_2}}$ are the Fresnel Kirchhoff Diffraction parameters, $\phi_{dif,front}$ and $\phi_{dif,back}$ represent the phases of diffraction signal from each side, and d_1 and d_2 means the vertical distance from TX and RX to the visitor's walking path, respectively [51]. The target crossing the FFZ at a uniform and relatively low speed will cause a weakening of the direct signal, on the one hand, and the strengthening of the diffraction signal, on the other hand, as Figure 2(c) depicted. Although the fluctuation durations of different subcarriers induced by human activities were similar, the specific changes were subtly different, thus distinct subcarriers can be used as the basis for further feature mining.

3.3 Challenges and Verifications

This section discusses the challenges we need to tackle before it works:

- **Finding uniqueness of head-count patterns from raw CSI.** A motivation experiment was conducted with two volunteers. We set the case of TX on the right side of the volunteer as the direction In and the opposite direction as Out. Although the identification can be achieved using a template-based dynamic programming approach, not only were there outliers in the amplitudes, but it was difficult to distinguish small differences, e.g., 1 people and two people in the case of Figures 3(a) and 3(b). Visitors' physical traits can be a hindrance in calculating the head count, but they can also be used to discern demographic segmentations.



(a) Single target in. (b) 2 targets in. (c) Single target out.

Fig. 3. The amplitude of subcarrier 1 in different situations.

- **Looking for direction-related clues in the signal.** Head-count identification is useless if the system cannot judge the flow's direction correctly. Although the amplitude of a single RX should be symmetrical based on the Fresnel zone model, the actual received signal for a human model crossing the zone is asymmetric [51] due to the asymmetrical front and back of the body. However, no matter which direction a person passes, it is always the forehead that approaches the link first, so a single RX antenna cannot theoretically identify the direction between Figures 3(a) and 3(c). Even though the asymmetry of the antenna orientation map and the environment may cause subtle nuances in direction, it is considered that they are not stable in all cases. When multiple users move in and out simultaneously, the problem becomes more complicated.
- **Reducing training time costs when new situations are introduced.** Each fixed-location device may face the possibility of adding new samples later to increase the diversity of identification, which significantly adds the time cost. Specifically, if the system is trained to recognize up to three people and later wants to increase to four people, then the network structure has to be adjusted and fully retrained.

4 SYSTEM DESIGN

4.1 Overview

In response to the challenges of Section 3.3, the architecture of the proposed system consists of six components (see Figure 4):

- **Data collection:** TX is set to transmit packets at a high frequency to RXs, and the raw CSI data from two RX antennas (abbreviated as RX1&2) is leveraged.
- **Signal pre-processing:** The Hampel filter is utilized to remove outliers and the **Discrete Wavelet Transform (DWT)** is used to roughly extract the main features of the amplitudes. Afterward, the variance is calculated in real time with a sliding window to segment the samples when visitors pass instantly. The segmentation of RX1 is applied directly to RX2 to ensure the asynchronous of the two antennas in time.
- **Crossing head-count/direction/visitor-type detecting:** We use the deep-learning method to enhance the robustness of the system without feature engineering. After pre-processing the samples, a two-channel CNN model is presented to further extract the features and classify the head count and direction. Since different body shapes mask the signal differently, the model is expected to learn about the visitor types.
- **Upgrade module:** The parameters of a trained model can be reused for training in new situations with transfer learning.

4.2 Signal Pre-processing

To overcome the interference caused by the environment and hardware, the input signals are first passed through the Hampel filter[1] to remove outliers. Figure 5(b) illustrates the outcome of the Hampel function with Figure 5(a)

[1] https://ww2.mathworks.cn/help/signal/ref/hampel.html.

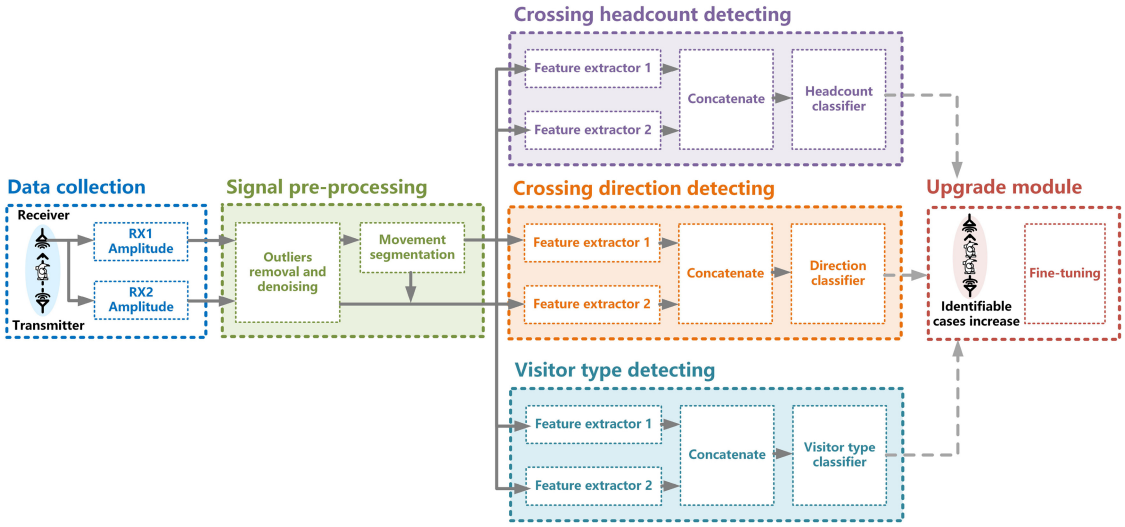
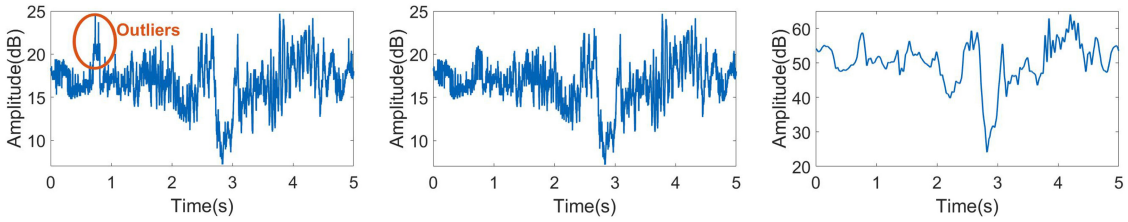


Fig. 4. System overview.



(a) Original amplitude. (b) Amplitude after Hampel filter. (c) Amplitude after DWT.

Fig. 5. Denoising the amplitude of subcarrier 1 when a target passing.

as input. It can be seen that the outliers are eliminated and the diffraction part remains intact. The generic step of DWT splits the signal into an approximation coefficient vector and a detail coefficient vector [1]. Since the waveform of the signal is mainly represented by the low-frequency part, DWT is leveraged to reduce the signal on multiple scales and obtain the approximation coefficients. Specifically, the signal is decomposed to six levels with the Daubechies 3 wavelet, which performed well in eliminating the residual noise from the fingerprint-induced sonic effect [57]. The system reconstructs the approximation coefficients for levels 4–6, respectively, from the decomposition structure, and then the three-layer approximation coefficients are added and smoothed, as Figure 5(c) shows. The processed amplitudes in different cases illustrate the basis for identifying the number, direction, and type of people.

As the signals obtained in practice are continuous, the automatic cutting of samples with visitor traffic is necessary for real-time detection. As illustrated in Figure 6, a sliding window moves from left to right by the sample point on the sub-figure above, and the corresponding variance change is shown below. The movement of people causes the growth of variance, and the superposition of DWT on the amplitudes helps to reinforce this difference. A proper threshold value should be selected, based on which it is possible to find the corresponding approximate time region of each sample. We record the first and last sample point above the threshold in each sample, and a length of the time window is added on both sides of the region to ensure that the target is fully captured.

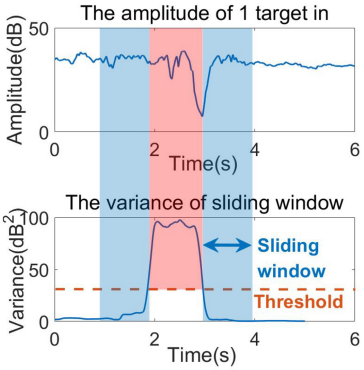


Fig. 6. Movement segmentation with a sliding window.

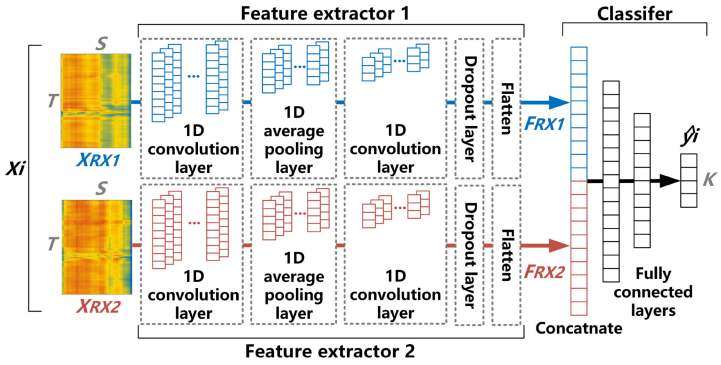


Fig. 7. The CNN structure.

4.3 Crossing Head-count Detecting

After getting the cut signal, the challenge is to ascertain the size of flow it corresponds to, which specifically means constructing the unique relationship between visitor traffic and CSI amplitudes. Deep learning algorithms are used to track the delicate head-count-related variables, because the data received from commercial WiFi equipment is highly unstable due to transceiver asynchrony, hardware noise, and rich multipath [23]. CNN is a network that can automatically extract local spatial information from a large number of training samples, abstract them into high-dimensional features, and perform classification. A CNN structure generally includes convolution layers, pooling layers, dropout layers, flatten layers, fully connected layers, and so on. Depending on the moving direction of the kernel, it can be classified into 1D, 2D, and 3D types. The first two types of CNNs are commonly used in signal classification problems, for the signal contains both time-domain features and frequency-domain features. Specifically, mapping a 1D signal to the frequency domain by **short-time Fourier transform (STFT)** results in a 2D time-frequency distribution map. Some studies used these spectrograms for 2D-CNN identification [84, 91]. Although there are studies that put the pre-processed feature matrixes into the 2D neural network because of the similarity in data structure [29, 63], 1D-CNN is compatible with array signal processing [39, 48], leading to smaller computational cost and higher performance in simple classification [44]. Therefore, inspired by Reference [86], a two-channel 1D-CNN model is proposed as Figure 7. It first extracts the features from each of the two antennas independently using two separate CNN models and then merges them for classification. Before the samples are put into training, linear interpolation is leveraged to keep the size uniform. The sample size after interpolation is $T \times S \times 2$ (see Figure 7), which can be expressed as

$$X_i = [X_{RX1} = f_{pre}(abs(H_{(1,1)})), X_{RX2} = f_{pre}(abs(H_{(1,2)}))], \tag{6}$$

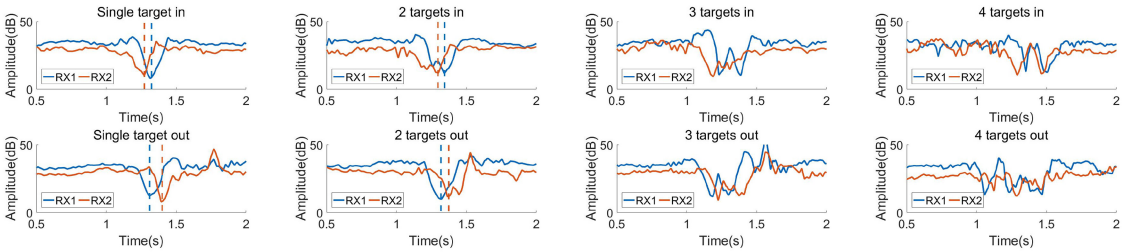
where T means the specified length of time, S means the number of subcarriers, and $f_{pre}(\cdot)$ means the pre-processing and interpolation of the amplitudes, respectively. For each X_i there is a corresponding y_i as a label. The label denotes the real head count. The samples of two RXs are fed in two independent feature extractors, respectively. The output of feature extractors serves as the deeper features:

$$F_{RX1} = f_{FE1}(X_{RX1}, \theta_{FE1}), F_{RX2} = f_{FE2}(X_{RX2}, \theta_{FE2}), \tag{7}$$

where θ_{FE1} and θ_{FE2} mean the parameters in each feature extractor. The two deep features are concatenated for classifier:

$$\widehat{y_i} = f_C(F_{RX1} \oplus F_{RX2}, \theta_C), \tag{8}$$

where θ_C means the parameters in the classifier, and $\widehat{y_i}$ means the prediction of the model based on X_i, θ_{FE1}, θ_{FE2}, and θ_C. The length of y_i and $\widehat{y_i}$ is the max number of identifiable concurrent people during data collection. y_i is



(a) Single target in and out. (b) 2 targets in and out. (c) 3 targets in and out. (d) 4 targets in and out.

Fig. 8. The denoised amplitudes of subcarrier 1 in RX1 and RX2 when different number of targets passing in the opposite directions.

the one-hot encoding vector of the category and the subscript of the maximum value in $\widehat{y_i}$ means the predicted result. The categorical cross-entropy loss is calculated to reflect the accuracy of the forecast:

$$Loss(\theta_{FE1}, \theta_{FE2}, \theta_C) = -\sum_{n=0}^{|X|}\sum_{k=0}^{K} y_{(i,k)} log(\widehat{y}_{(i,k)}), \tag{9}$$

where $|X|$ means the total number of the input samples and K means the number of corresponding categories, equal to the length of the predicted label. The purpose of iterative training of CNN can be expressed as

$$(\widehat{\theta}_{FE1}, \widehat{\theta}_{FE2}, \widehat{\theta}_C) = \underset{\theta_{FE1}, \theta_{FE2}, \theta_C}{\arg\min} \; Loss, \tag{10}$$

where $\widehat{\theta}_{FE1}$, $\widehat{\theta}_{FE2}$, and $\widehat{\theta}_C$ represent the updated network parameters. Ideally, iteration of the parameters minimizes the loss, so the predictions become increasingly accurate with training.

4.4 Crossing Direction Detecting

Following the estimation of visitor traffic, it needs to determine the direction of each crossing event for achieving accurate cross-regional head counting. The direction identification module faces the challenge of signal sensitivity similar to the previous section, and both parts can be executed in parallel. Figure 8 plots the processed amplitudes of two RXs on a receiver when different numbers of targets cross the LoS vertically. As shown in Figure 9, since the line of the two RXs and the direction of visitor routes are parallel, even the RXs are close to each other, changes in one RX antenna generally precede the other in time, which remains relatively stable in the case of up to four people. Therefore, the order in which the fluctuations of the two antenna signals occur may give more basis for the direction identification beyond the effects of environmental and hardware asymmetry. It is thus worth noting that the RX1 signal segmentation findings must be applied directly to the RX2 signal to ensure the time difference between the two antennas, rather than splitting the two antennas separately. Simultaneously, the activity of numerous targets increases the complexity of amplitude fluctuations, rendering manually derived features such as wave-valley time disparities invalid [see Figures 8(c) and 8(d)]. As a result, in theory, the identical structure of the two-channel CNN can be used for direction determination. In the network used to identify the direction, y_i and $\widehat{y}_i$ denote the true and predicted label of direction, respectively (In and Out), and both are one-dimensional vectors of length 2 ($K = 2$).

4.5 Visitor Type Detecting

The importance of segmented visitor research keeps rising, and demographic segmentation is a classic classification principle for museum visitors. While Falk [22] argued that demographic variables alone may not provide the best explanation for museum service enhancement strategies, targeting segments such as children exceptionally

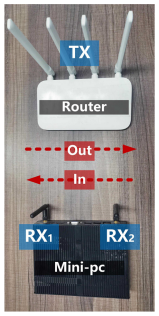


Fig. 9. The devices in experiments.

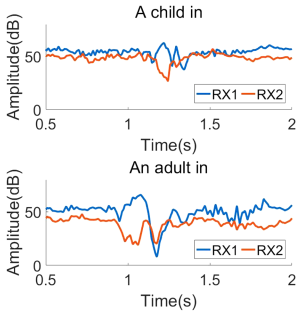


Fig. 10. Comparison of signal changes induced by an adult and a child.

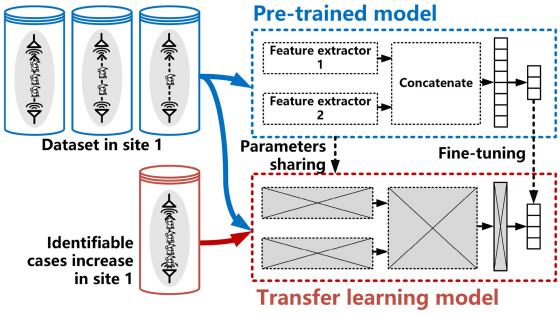


Fig. 11. Transfer learning when identifiable cases increase.

seems to be of greater practical relevance [2]. Non-visual-based wireless sensing methods ensure privacy but also somewhat lose the opportunity to learn more about the characteristics of the audience. However, it is still possible for the signal changes to describe the approximate body state. [51] compared the amplitude changes of the barrel and the dummy model across the link. As in Figure 10, our results also show the difference between a child and an adult when crossing the LoS. Therefore, with similar model training, the system can classify the types of visitor traffic in more detail. The module was examined in Experiment 2.

4.6 Transfer Learning

The ideas of transfer learning are proposed to overcome the problem of machine learning migration between different knowledge domains, one of which is pre-training and fine-tuning. It has been successfully applied in device-free user identification to reduce the retraining time when the number of users grows [34]. We use a similar approach to verify its effectiveness in flow counting. Since the procedures of shallow feature extraction may be comparable, if a trained CNN model works well in one configuration, portions of the model's parameters can be frozen and reused for training in others. The partial fixation of parameters results in a reduction in training time. The process of transfer learning can be presented as

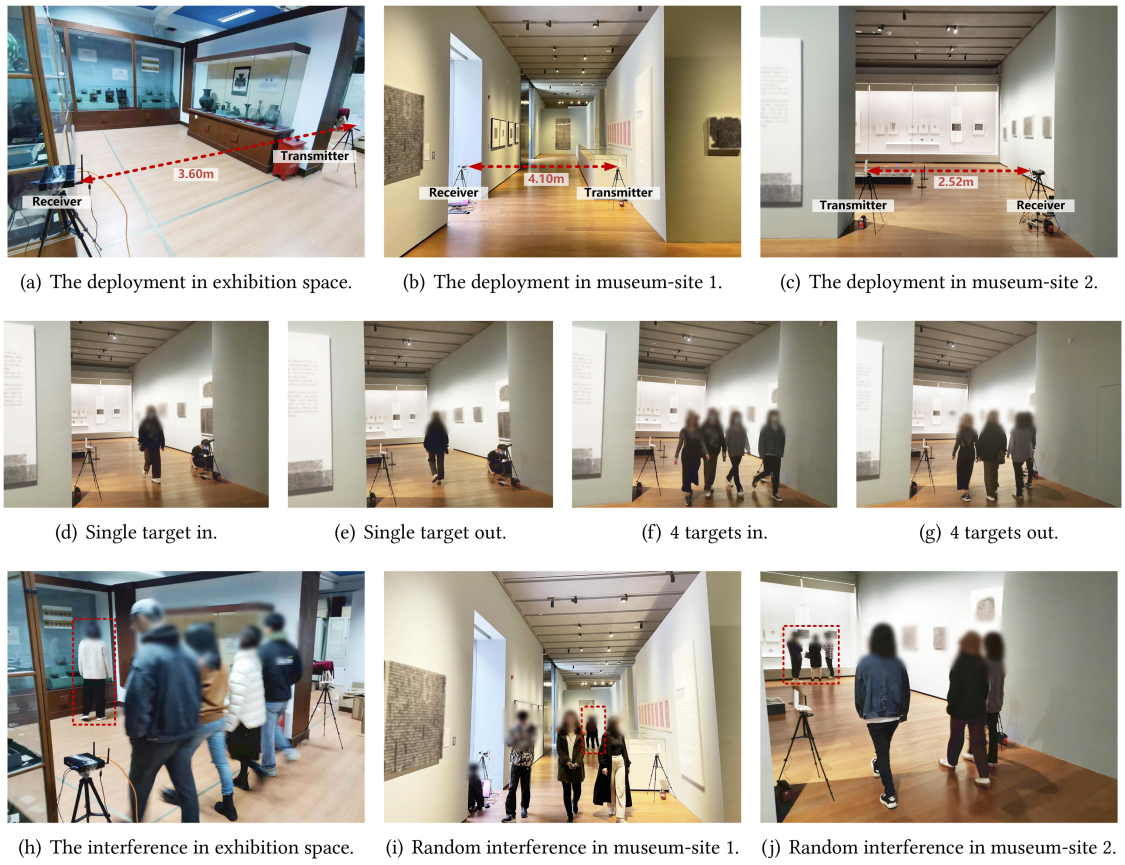
$$(\widehat{\theta_{trainable}}) = \underset{\theta_{trainable}}{\arg\min}\ Loss, \tag{11}$$

where $\theta_{trainable}$ represents the updateable parameters in pre-trained CNN. Figure 11 shows how pre-training and fine-tuning are applied in the upgrade module. Assume there is a trained CNN model adequate for a three-person bi-directional head counting at a given place with a particular number of samples. When the recognition capacity has to be increased to four people, the old data set and the new samples can be mixed for retraining. In this study, we trained the transfer learning model on the modified final fully connected layer. Existing feature extractor parameters are fixed, and only the classifier's exit number needs to be increased. By fine-tuning the last layer of a pre-trained CNN, it avoids retraining the entire network, reducing training time and boosting parameter reusability.

5 EXPERIMENT 1

5.1 Experimental Setup

We used a Xiaomi AC1200 router as the transmitter and a mini-pc equipped with Intel 5300 WiFi NIC as the receiver. The router operated on the 5 GHz frequency range, with a distance of 0.15 m between the two RXs. According to References [84, 91], a high sampling rate generally enables an increase in accuracy, therefore the mini-pc was set to receive CSI packages every 0.002 s. Three independent experiments were conducted in the



(a) The deployment in exhibition space. (b) The deployment in museum-site 1. (c) The deployment in museum-site 2.

(d) Single target in. (e) Single target out. (f) 4 targets in. (g) 4 targets out.

(h) The interference in exhibition space. (i) Random interference in museum-site 1. (j) Random interference in museum-site 2.

Fig. 12. Different scenarios with different deployments in Experiment 1.

college's heritage exhibition space and two sites in the Zhejiang University Museum of Art and Archaeology, respectively. The exhibition space was mainly used for artifact storage, facilitating strict control of variables. To minimize the impact on other visitors, the museum's field experiments were held on weekdays. As the testbeds shown in Figures 12(a)–12(c), different deployments were applied according to the features of the three environments. The distance between transmitter and receiver was 3.60 m in the exhibition space, 4.10m in site 1, and 2.52 m in site 2 of the museum. According to Equation (3), the short axis of FFZ was 0.23 m, 0.25 m, and 0.19 m in the three environments. In consideration of the possibility of the younger audience, the height of the two devices was fixed at 0.85 m.

Ten volunteers composed of seven females and three males were invited to participate in the experiment. The volunteers were between 1.55 m and 1.77 m in height. A total of 1,080 samples were collected, as summarized in Table 2. According to our preliminary observation at the museum, no more than three people generally pass through the regional border at the same time, thus the data for four types of head count × In/Out = 8 categories was collected. Data in the three sites were sampled on three distinct days, and the volunteers were not exactly repeated each time. The experimental scenarios for different cases are shown in Figures 12(d)–12(g). The volunteers were only asked to pass through the LoS in the required direction as much as possible at the same time with other companions. A researcher was in the vicinity of the receiver, delivering instructions at the start and end of each sample collection. There were no further prerequisites for volunteers aside from those specified above.

Table 2. Summary of Data Collection in Experiment 1

Experimental site	Participants	One person		Two people		Three people		Four people		Total
		In	Out	In	Out	In	Out	In	Out	
Exhibition space	Six volunteers	40	40	48	48	48	48	48	48	368
Museum-site1	Four volunteers	40	40	48	48	48	48	48	48	368
Museum-site2	Four volunteers	40	40	48	48	48	48	36	36	344

They were allowed to talk, watch their smartphones, and freely determine their speed, stride, and arm swing during the walk.

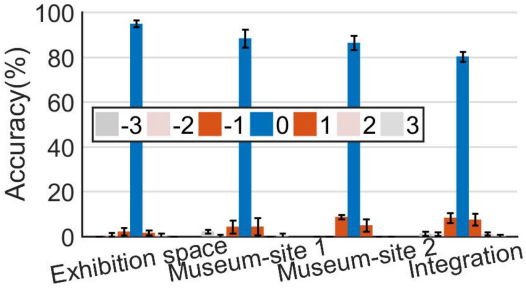
In the exhibition space, we chose a volunteer at random to act as an extra audience for half of the samples, simulating possible interference in real life. As long as the interferer did not cross the LoS, he was free to walk or stop in the experimental area at any time during the experiment. To further ensure the diversity of the samples, one-, two-, three-, and four-people cases were, respectively, derived from 8, 12, 24, and 48 different combinations (such as changing the selection of targets and interferer in different rounds, altering the order of relative positions, etc.). In the museum, we still collected the volunteers' behavioral data out of concern for other audiences' experiences, but there were also random disturbances from visitors around the experimental environment. Figures 12(h)–12(j) illustrates the potential interference that was deliberately designed or randomly appeared in the experiments. Similarly, in site 1 the four cases were derived from 4, 4, 12, and 24 different combinations, and for site 2 there were 4, 4, 12, 18 ways to combine, respectively.

In terms of movement segmentation, suitable variance thresholds were confirmed according to the data in different environments. Only 2 of over 1,000 samples were not detected. Since the undetected samples are also real data, they were directly adopted as samples without segmentation. Each RX sample included 30 subcarriers, and the standard duration of linear interpolation was set to 1,000, which is equivalent to 2 s for the experimental sampling rate. The CSI streams were extracted and pre-processed with MATLAB R2016a, and the CNN model was constructed and trained based on Tensorflow and Keras 2.7.0.

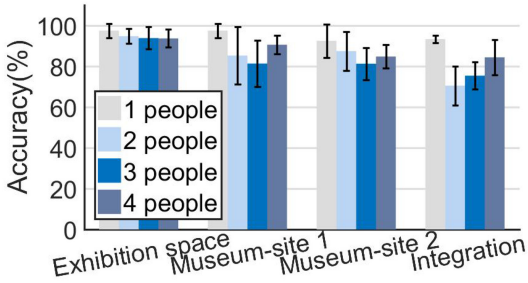
5.2 Results

Overall performance. K-fold cross-validation was performed to assess the validity of the method. The value of K was set to 5, which means the data set was randomly divided into five mutually exclusive parts. In turn, four of them were used as a training set and the remaining part was used as a testing set. The best accuracy of the trained model on the testing set was recorded for each round. After five repetitions, the mean and standard deviation of the accuracies were calculated as a measure. The CNN parameters were set as follows: batch size = 10, learning rate = 0.001, Conv1D filters = 64, kernel size = 10, pool size = 2, dropout rate = 0.2, dense units = 256, 128. For each site, each sample was normalized with the maximum and minimum values of training data as intervals. Figure 13(a) illustrates the proportion of test samples with accurate and incorrect predictions in different environments. In the exhibition space, the average recognition accuracy for head-count detection was 94.84%. The accuracy of counting dropped to 88.32% and 86.33% in the museum's 2 ground situations, respectively. However, approximately 74.43% and 100% of the forecast errors were only 1 person away from the ground truth. The maximum standard deviation of cross-validation for all sites was 4.05%. It demonstrates that the proposed strategy worked in a variety of museum environments/deployments and that neither the time of collection nor the presence of other visitors in the area had a significant impact on the outcomes. From Figure 13(b), the model rarely missed the determination of 1 target, but the two-, three-, and four-target cases were more likely to cause confusion. Integrating all samples for cross-validation, the accuracy of head-count identification reached 80.27%, which shows that the environmental differences affected the generalizability of feature mining (for more discussion, see Section 8.1).

Furthermore, the average accuracy of direction recognition in the 3 environments was 99.73%, 91.84%, and 99.13%, respectively, with 95.09% for integration. The maximum standard deviation of cross-validation for all



(a) Proportion of prediction error.



(b) Prediction accuracy for different number of people.

Fig. 13. The recognition accuracy of head count.

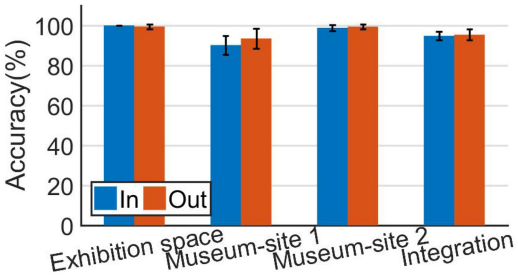


Fig. 14. The recognition accuracy of direction.

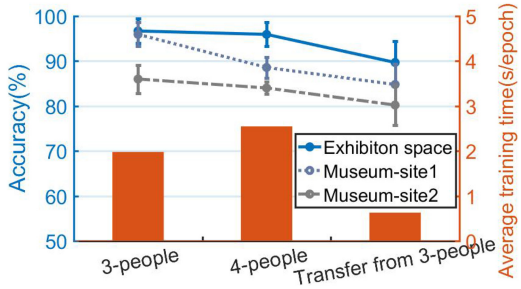


Fig. 15. Performance of transfer learning in three environments.

sites was 2.38%. Figure 14 showed the specific performance, and the accuracies were similar in both directions. The proposed method yielded relatively low accuracy in direction detection at museum-site 1, most likely due to the multipath effect produced by the router's closeness to the whole wall as Figure 12(b).

Transfer learning when the identifiable maximum number of people increases. Following the assumption in Section 4.6, the migration performance of the three or four people was verified for each of the three settings. For each environment, the models for the three-person case with cross-validation were saved. The output of the pre-trained three-person model's penultimate layer was spliced into a modified classifier with 4 exits. The original training set and the fourth person's training samples were used to update the weights of the new model. To expedite convergence, the learning rate was set to 0.01. Fivefold cross-validation accuracy and the average training time per epoch are shown in Figure 15. Overall, although the effectiveness of fine-tuning was limited by the original model's accuracy, the transferred model retained an accuracy decrease of 6.25%, 4.28%, and 4.52% in the three sites while saving around 75.31% of the average training time. Taking exhibition space as an example, for the same training/testing set division on the data of four people, the model with random initialization achieved an average accuracy of 95.94% at an average time cost of 2.53 s/epoch. However, if the pre-trained three-person model with an average accuracy of 96.69% was partly used for initialization, the average accuracy was maintained at nearly 90% while requiring only 0.64 s/epoch.

6 EXPERIMENT 2

Based on the implementation of the elementary functionality, the goal of Experiment 2 is to evaluate the system's effectiveness in recognizing visitor types. The science museum is a typical example, where parents and children serve as the main visitor groups. It is expected that variation in the physical characteristics of adults, children,

(a) The deployment in Experiment 2. (b) A child out. (c) An adult-child group out.

Fig. 16. The scenario and deployment in Experiment 2.

and adult-child groups may differently affect the fluctuations of the signal and thus function as a cue for visitor type identification.

6.1 Experimental Setup

We selected an open space in front of an interactive installation that was closed to the public due to the epidemic in the Zhejiang Science and Technology Museum, and, unlike Experiment 1, we recruited parent-child volunteers who visited the site as volunteers. The distance of the transmitter-receiver pair was 2.73 m (the short axis of FFZ was 0.20 m). Prior to the start of the experiment, the researchers explained the experimental design, the content of the data collected, the safety of the signal, and the ethical norms of the study to the interested visitors in the form of posters and oral presentations, inviting them to participate in one of the three experimental levels (adult, child, and adult-child group). Each experiment consisted of 15 round trips in and out of LoS and lasted about 10 minutes (see Figure 16). The relative positions of the adult and child in the group were also noted for balance when sampling. The sample size may be reduced depending on the interest of the children, and all the volunteers received a souvenir after the experiment. A total of 6 adults, 10 children, and 10 additional adult-child pairs participated in the data collection, and their height and weight are shown in Figure 17. Adults' age ranged from 24 to 59 years old, and children's age ranged from 2 to 9 years old. Because of the age constraints at the time of subject recruitment, there were more substantial disparities between the body postures of adults and children. Aside from that, there were minor variances between groups and singles, and the Body Mass Index among adults and children ensured diversity. In addition, 55 free samples were collected when no one was passing by, and a total of 816 samples were collected as listed in Table 3.

Experimentally, it was found that relying solely on detecting the variance value of the sliding window tends to cause incorrect segmentation. The selected threshold missed only one child sample while incorrectly detecting 10 unaccompanied samples. It was possible that other factors also caused large fluctuations. Therefore, the unidentified samples (including the free ones) were linearly interpolated and then fed into the CNN training. The experimental setups except for the above description were the same as Experiment 1.

6.2 Results

Overall performance. Visitor type classification training was performed using fivefold cross-validation for free, child, adult, and group cases. The mean value of five times recognition accuracy was 88.97%, the standard deviation was 3.59%, and the confusion matrix was summed as shown in Figure 18. Unlike laboratory experiments, the volunteers, especially children, did not always follow a similar walk each time when sampling, which created a challenge for machine learning with small samples. Overall, a number of free samples were still recognized as passages, and the adult-child groups were easily confused with the single-person samples.

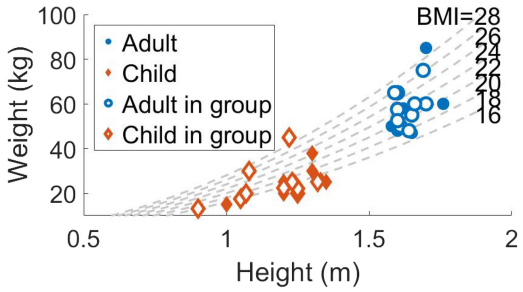


Fig. 17. Physical data of the volunteers in Experiment 2.

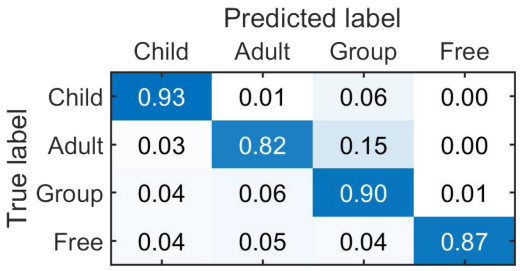


Fig. 18. The confusion matrix of visitor type classification.

Table 3. Summary of Data Collection in Experiment 2

Visitor type	Participants	In	Out	Total
Child	10 volunteers	145	146	291
Adult	6 volunteers	90	90	180
Adult-child group	20 volunteers	145	145	290
Free	—	55		55

The mean value of recognition accuracy of direction detection was 91.46%, and the standard deviation was 1.93%, of which the average values of In and Out identification accuracy were 90.81% and 92.11%, respectively.

7 EXPERIMENT 3

Due to the museum's broad aisle spacing and diverse exhibition routes, simultaneous entry and exit is highly likely, which has not been extensively examined in relevant studies. Experiment 3 was conducted to demonstrate if CSI data could differentiate more complex circumstances. When utilized for regional head counting, some cases can be merged when numerous persons enter and exit at the same moment, and the upgrade module can also add these scenarios to the system.

7.1 Experimental Setup

Experiment 3 recruited three volunteers to conduct data collection in an empty art gallery of the school. The distance of the transmitter-receiver pair was 2.98 m (the short axis of FFZ was 0.21 m). All volunteers were female university students, 1.58 m, 1.60 m, and 1.65 m in height, respectively. The most complex combination in the three-person case and the factor of speed were considered, some of which are shown in Figure 19. Before the experiment began, the researcher measured the volunteers' gait speed in self-perceived fast and slow situations and obtained the average speed of 1.178 m/s and 0.78 m/s, respectively. The volunteers followed a similar pace in the formal experiment. Taking into account the simultaneous entry and exit, there are two, four, and eight cases of one- to three -person combinations, and for multi-area visitor counting, we further categorized them into the nine cases as in Table 4. Fifteen samples were collected in each level for the single target and 18 samples in each level for the multiple targets case, for a total of 492 samples. The relative positions of multiple targets were balanced. The experimental setups except for the above description were the same as Experiment 1. The selected variance threshold missed a sample of the single target, which was also linearly interpolated and put into the dataset.

7.2 Results

Overall performance. Fivefold cross-validation was likewise used to evaluate the nine-case classification. The classification mean accuracy and standard deviation were 88.21% and 3.42%, respectively. The sum of the

(a) 2 targets cross, combination 1. (b) 3 targets cross, combination 1. (c) 3 targets cross, combination 2.

Fig. 19. The scenario and deployment in Experiment 3.

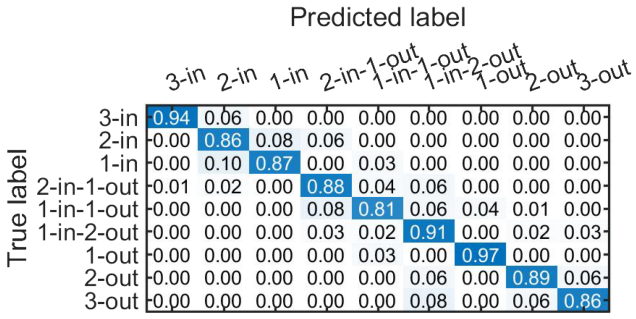


Fig. 20. The confusion matrix of classification in Experiment 3.

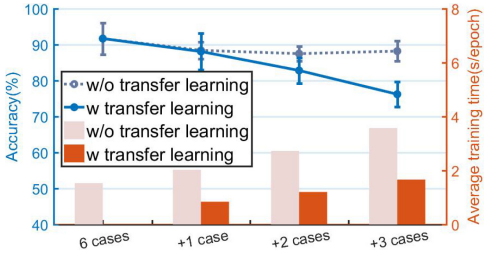


Fig. 21. Transfer learning in Experiment 3.

Table 4. Summary of Data Collection in Experiment 3

Speed	1-in	1-out	2-in	2-out	1-in-1-out	3-in	3-out	1-in-2-out	2-in-1-out	Total
Fast	15	15	18	18	18×2	18	18	18×3	18×3	246
Slow	15	15	18	18	18×2	18	18	18×3	18×3	246

testing set confusion matrices is shown in Figure 20. Among the samples with wrong predictions, 70.69% had a deviation of one person and 29.31% had a deviation of two persons.

Transfer learning when the identifiable maximum number of cases increases. To evaluate the optimization effect of the upgrade module, the originally proposed model was first trained with three users In and Out for a total of six cases, and the average training time and accuracy were measured as a baseline. The parameters of the model were frozen and the last classifier was structurally modified and trained for 6+c cases ($c \in [1,3]$). As shown in Figure 21, similar to the results of Experiment 1, transfer learning significantly reduced the training time. However, the compatibility capability of the original model was limited when the number of updated cases was too high. When c = 3, although fine-tuning saved around 53.44% of the time, it came at the cost of over 10% average decrease in accuracy.

Comparison of different signal features. To further discuss the possibility of simplifying the device, the efficiency of a single antenna was examined. Besides, since the amplitude ratio of the array antennas in hardware remained constant in the unoccupied state, the features of the two antennas can be combined during preprocessing. The ratio of the two RXs' absolute values was found in the segmented area, afterward, the outlier removal, hop elimination, and DWT were performed sequentially. We thus compared the proposed method with a single-input CNN of other features in the five sites using fivefold cross-validation (see Figure 22). The results

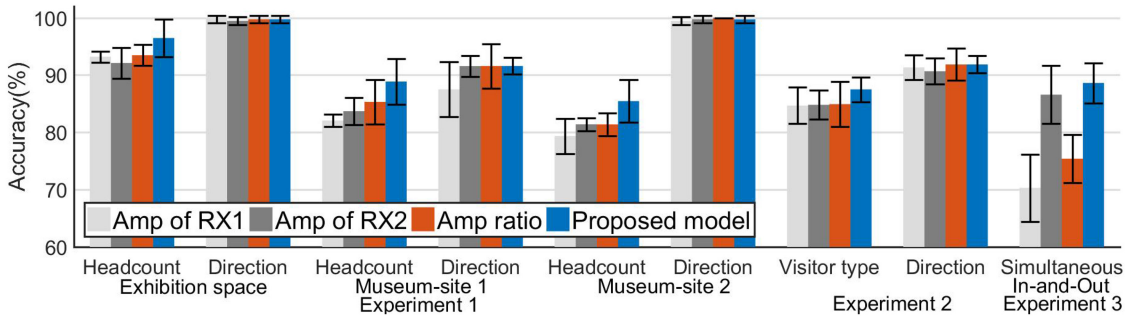


Fig. 22. Comparison of performance with different signal features.

reveal that a single antenna could provide astonishing accuracy, but that it was not stable in all experiments. The amplitude ratio had some advantages in direction identification, however, the instabilities of one antenna may impair feature engineering (as Experiment 3). Overall, the proposed approach retained a relative advantage in most cases, though the findings show that there is still room for simplified or expanded deployment options.

8 THE POTENTIAL AND LIMITATIONS OF CONTACTLESS SENSING IN MUSEUMS

There is much more to consider in a real museum setting. The problems of intensive labor, the effects of front-to-back distance, and more potential applications of device-free sensing were discussed, providing an outlook for future work on the existing problems.

8.1 Cross-domain Detecting

Intensive labor is one of the impediments to wireless contactless sensing in smart space applications. Machine learning methods often require a large number of samples, and since museums are social public spaces, it is more difficult to collect data for training. This requires the model to exclude environmental or personal factors from the feature vector as much as possible to ensure that the system trained in the source domain can be directly applied to other target domains. In our experiments, even though the samples in each environment have been normalized independently, it was difficult for a pre-trained model to guarantee recognition across environments, even for the most basic binary classification. Some signal fluctuations were discovered to differ from the theoretical ones in Figure 8 to a certain extent, probably due to the proximity of the router to the wall, excessive masking of the FFZ, or other personal, environmental, and hardware influences. Although pre-training and fine-tuning can save time cost, manual labeling and sampling for each monitoring site are difficult to avoid.

To reduce the workload of manual labeling, **Domain Adaption (DA)** [24] and its variants [9, 15, 36] have been involved in minimizing environmental factors in signal abstraction. DA learning adds a discriminator to judge the domain of the deep feature from the feature extractor. In the training phase, the source domain samples with labels and the target domain samples without labels enter the feature extractor, and only the feature vectors of source domain samples enter the classifier, while those from both domains enter the domain discriminator. The optimization goal is to minimize the classifier loss and maximize the discriminator loss so that the feature extractor performs basic classification while deceiving the discriminator as much as possible to achieve migration across environments without manual labeling of the target domain samples. We selected 2 environments with relatively similar signal characteristics and used the DA model to implement transfer learning for direction recognition from museum-site 2 toward the exhibition space. The code was adapted from https://github.com/Daipuwei/DANN-MNIST, where we replaced the network layers with a structure similar to the one in this article. As Figure 23 shows, the feature extractors fused data from both environments and achieved a direction recognition accuracy of 93.33% for unlabeled target domain samples. However, between

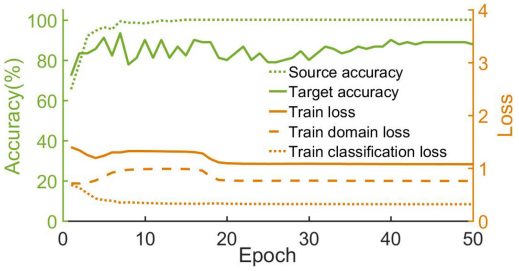


Fig. 23. Domain adaption learning from museum-site 2 to exhibition space.

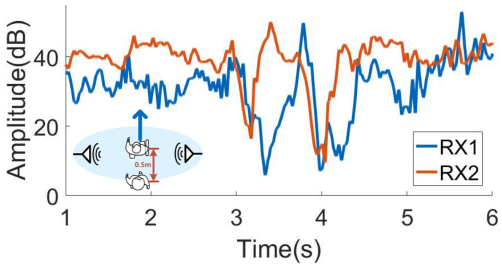


Fig. 24. The denoised amplitudes of subcarrier 1 in RX1 and RX2 when two targets passing with a 0.5 m front-to-back distance.

datasets from other sites, adversarial training played a minor role, which could be attributed to large differences in surroundings and volunteers, or insufficient manual signal feature extraction. Moreover, DA still necessitates a sizable number of samples from the target domain.

To further reduce manual sampling costs, the most efficient method is to extract stable features that do not require extra training. Zheng et al. calculated the domain-independent feature for zero-effort cross-domain gesture recognition through multiple wireless links [90]. Data augmentation is another common method to cope with data scarcity. Um et al. proposed various data augmentation methods for wearable sensor data and applied them to enhance the classification performance of motor states for Parkinson's disease patients [72]. **Generative Adversarial Network (GAN)** allows the generator to produce simulations close to the real data, thus increasing the number and breadth of signal samples [3, 7]. The meta-learning framework could also be used to mitigate excessive training data collection. In machine learning, meta-learning tries to create models that can swiftly adjust to new scenarios that have never been trained before, allowing for few-shot activity and speech recognition adaptation [16, 26].

8.2 The Impact of Front-to-back Distance

Since in practice there may be a small front-to-back spacing between two people passing through the link, we must determine the maximum front-to-back distance at which the method works. An experiment was conducted in an empty office with a 2.0 m distance between the transmitter and receiver. Two volunteers held both ends of the tapeline to stabilize the relative distance. Figure 24 reports the processed amplitudes caused when the front-to-back distance between the two people was 0.5 m. Even though the visitors were separated by 0.5 m, the features of the two antennas in direction were still kept. Considering the effect of front-to-back spacing, the proposed system requires a balance between the length of the sliding window used for segmentation and the addition of new cases.

8.3 Other Possible Applications of Contactless Sensing in Museums

Experiment 3 also tried a rough binary classification of walking speed. The mean value of recognition accuracy of speed classification was 87.59%, and the standard deviation was 2.68%, of which the average values of fast and slow identification accuracy were 89.44% and 85.75%, respectively. For a single target, calculating the time difference that causes fluctuations at the edge of the FFZ may help to measure the velocity. Although the trait of walking speed-induced signal changes for multiple targets is difficult to interpret, the result provides preliminary evidence that the presented method may be useful for assessing museum fatigue or detecting the elders.

Another future work is to validate and improve more applications of the WiFi-based contactless systems in the museum scene. The same commodity WiFi devices worked well in activity identification [77, 78], gesture recognition [45] and route tracking [54], and so on. According to embodied cognition theory, in addition to

body position, visitors' gestures may also be variables in understanding their appreciation of art and social processes [68]. There have been studies that extend the fine-grained detection of a single person to several targets [73, 88]. These studies on the prevalence of WiFi perception are expected to promote museum services and visitor research.

In addition to the discussions above, there are some other inadequacies in the study. First, the relatively small sample size is insufficient to adequately explain the effects of more variables including spatial layout, WiFi wavelength and TX frequency, walking angles and finer-grained speeds, items such as backpacks carried by visitors, and so on. The identification of some of these variables may provide additional low-cost and efficient insights for museum studies. The precise distinction between the occupied and unoccupied states is also not thoroughly discussed. Furthermore, because there is no fixed relationship between the complexity of the CNN structure and its recognition efficiency, the model design may need to be tweaked as training data grows more complex. Due to a lack of model interpretability, it is unknown to which extent the signal features of the demonstration were utilized in the machine learning session.

9 CONCLUSION

This article presents and tests a WiFi-based contactless multi-area visitor counting system that adapts to the museum environment, illustrating the potential and limitations of contactless sensing methods applied to the visitors' time-space distribution behavior tracking in cultural spaces. Extensive experimental results show that the amplitudes of two receiving antennas can be effectively related to the size, direction, and type of visitor traffic with a two-channel CNN, and have the potential to be used for discriminating more complex situations to some extent. In comparison to previous research on WiFi-based contactless flow counting and direction recognition, our work considers compatibility with visitor studies and therefore delves further into device flexibility and improves scalability through transfer learning, while exploring the identification of visitor types as well as simultaneous entry and exit. The limitations of cross-domain recognition and the front-to-back distance of visitors are discussed with possible solutions. Adding samples, improving the robustness of different environments/deployments, and more effective detecting methods are still needed for the eventual goal of realizing it in museums.

REFERENCES

[1] Heba Abdelnasser, Moustafa Youssef, and Khaled A. Harras. 2015. WiGest: A ubiquitous WiFi-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'15)*. IEEE, 1472–1480. https://doi.org/10.1109/INFOCOM.2015.7218525

[2] Eeva-Katri Ahola and Liisa Uusitalo. 2008. *Can we Segment Art Museum Visitors? A Study of Segmentation based on Consumer Motives and Preferences*. Helsingin kauppakorkeakoulu, 157–168.

[3] Sudhanva Bhat and Enrique Hortal. 2021. GAN-based data augmentation for improving the classification of EEG signals. In *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA'21)*. ACM, 453–458. https://doi.org/10.1145/3453892.3461338

[4] Stephen Bitgood. 2009. Museum Fatigue: A Critical Review. *Visitor Studies* 12, 2 (Oct. 2009), 93–111. https://doi.org/10.1080/10645570903203406

[5] Ramon F. Brena, Juan Pablo García-Vázquez, Carlos E. Galván Tejada, David Muñoz-Rodriguez, Cesar Vargas-Rosales, and James Fangmeyer Jr. 2017. Evolution of Indoor Positioning Technologies: A Survey. *J. Sensors* 2017, Article 2630413 (Mar. 2017), 21 pages. https://doi.org/10.1155/2017/2630413

[6] R. Brunelli, O. Lanz, A. Santuari, and F. Tobia. 2007. Tracking Visitors in a Museum. In *PEACH—Intelligent Interfaces for Museum Visits*, Oliviero Stock and Massimo Zancanaro (Eds.). Springer, 205–225. https://doi.org/10.1007/3-540-68755-6_10

[7] Jiaxing Chang, Fei Hu, Huaxing Xu, Xiaobo Mao, Yuping Zhao, and Luqi Huang. 2021. Data augmentation of wrist pulse signal for traditional chinese medicine using Wasserstein GAN. In *Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences (ISAIMS'21)*. ACM, 426–430. https://doi.org/10.1145/3500931.3501003

[8] Lili Chen, Jie Xiong, Xiaojiang Chen, Sunghoon Ivan Lee, Daqing Zhang, Tao Yan, and Dingyi Fang. 2019. LungTrack: Towards Contactless and Zero Dead-Zone Respiration Monitoring with Commodity RFIDs. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 79 (Sept. 2019), 22 pages. https://doi.org/10.1145/3351237

[9] Xi Chen, Hang Li, Chenyi Zhou, Xue Liu, Di Wu, and Gregory Dudek. 2020. FiDo: Ubiquitous fine-grained WiFi-based localization for unlabelled users via domain adaptation. In *Proceedings of the Web Conference (WWW'20)*. ACM, 23–33. https://doi.org/10.1145/3366423.3380091

[10] Angelo Chianese and Francesco Piccialli. 2014. Designing a Smart Museum: When Cultural Heritage Joins IoT. In *Proceedings of the 8th International Conference on Next Generation Mobile Apps, Services and Technologies*. IEEE, 300–306. https://doi.org/10.1109/NGMAST.2014.21

[11] Jeong Woo Choi, Xuanjun Quan, and Sung Ho Cho. 2018. Bi-Directional Passing People Counting System Based on IR-UWB Radar Sensors. *IEEE Internet Things J.* 5, 2 (Apr. 2018), 512–522. https://doi.org/10.1109/JIOT.2017.2714181

[12] Andreas Christian. 2019. Participant Reactivity in an Exhibition: The Effect of Overt Observation on Engagement Times. *Visitor Studies* 22, 1 (Apr. 2019), 67–83. https://doi.org/10.1080/10645578.2019.1603739

[13] Saandeep Depatla and Yasamin Mostofi. 2018. Passive crowd speed estimation and head counting using WiFi. In *Proceedings of the 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON'18)*. IEEE, 1–9. https://doi.org/10.1109/SAHCN.2018.8397119

[14] Saandeep Depatla, Arjun Muralidharan, and Yasamin Mostofi. 2015. Occupancy Estimation Using Only WiFi Power Measurements. *IEEE J. Select. Areas Commun.* 33, 7 (July 2015), 1381–1393. https://doi.org/10.1109/JSAC.2015.2430272

[15] Cao Dian, Dong Wang, Qian Zhang, Run Zhao, and Yinggang Yu. 2020. Towards domain-independent complex and fine-grained gesture recognition with RFID. *Proc. ACM Hum.-Comput. Interact.* 4, Article 187 (Nov. 2020), 22 pages. https://doi.org/10.1145/3427315

[16] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. *RF-Net: A Unified Meta-Learning Framework for RF-Enabled One-Shot Human Activity Recognition*. ACM, 517–530. https://doi.org/10.1145/3384419.3430735

[17] Shing H. Doong. 2016. Spectral Human Flow Counting with RSSI in Wireless Sensor Networks. In *Proceedings of the International Conference on Distributed Computing in Sensor Systems (DCOSS'16)*. IEEE, 110–112. https://doi.org/10.1109/DCOSS.2016.33

[18] Shing H. Doong. 2018. Counting Human Flow with Deep Neural Network. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. 799–808. https://doi.org/10.24251/HICSS.2018.100

[19] Kira Eghbal-Azar and Thomas Widlok. 2013. Potentials and limitations of mobile eye tracking in visitor studies: Evidence from field research at two museum exhibitions in Germany. *Soc. Sci. Comput. Rev.* 31, 1 (Feb. 2013), 103–118. https://doi.org/10.1177/0894439312453565

[20] Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. 2020. Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics. In *Proceedings of the International Conference on Multimodal Interaction*. ACM, 107–116. https://doi.org/10.1145/3382507.3418890

[21] Pedro Escuer, Ana Mateo, Christopher McConnell, and John Schutes. 2014. *Refining Visitor Tracking for Museum Victoria*. Technical Report. Worcester Polytechnic Institute, Worcester, MA.

[22] John H. Falk. 2016. *Identity and the Museum Visitor Experience*. Routledge.

[23] Chao Feng, Jie Xiong, Liqiong Chang, Ju Wang, Xiaojiang Chen, Dingyi Fang, and Zhanyong Tang. 2019. WiMi: Target material identification with commodity WiFi devices. In *Proceedings of the IEEE 39th International Conference on Distributed Computing Systems (ICDCS'19)*. IEEE, 700–710. https://doi.org/10.1109/ICDCS.2019.00075

[24] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, Francis Bach and David Blei (Eds.). PMLR, 1180–1189. Retrieved from https://proceedings.mlr.press/v37/ganin15.html.

[25] Andrew B. Godbehere and Ken Goldberg. 2014. Algorithms for Visual Tracking of Visitors Under Variable-Lighting Conditions for a Responsive Audio Art Installation. In *Controls and Art: Inquiries at the Intersection of the Subjective and the Objective*, Amy LaViers and Magnus Egerstedt (Eds.). Springer, Cham, 181–204. https://doi.org/10.1007/978-3-319-03904-6_8

[26] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. MetaSense: Few-shot adaptation to untrained conditions in deep mobile sensing. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems (SenSys'19)*. ACM, 110–123. https://doi.org/10.1145/3356250.3360020

[27] Dimitris Grammenos, Giannis Drossis, and Xenophon Zabulis. 2014. Public Systems Supporting Noninstrumented Body-based Interaction. In *Playful User Interfaces*, Anton Nijholt (Ed.). Springer, Singapore, 25–45. https://doi.org/10.1007/978-981-4560-96-2_2

[28] D. Grammenos, X. Zabulis, D. Michel, P. Panteleris, T. Sarmis, G. Georgalis, P. Koutlemanis, K. Tzevanidis, A. A. Argyros, M. Sifakis, and C. Stephanidis. 2013. A Prototypical Interactive Exhibition for the Archaeological Museum of Thessaloniki. *Int. J. Heritage Digital Era* 2, 1 (Mar. 2013), 75–99. https://doi.org/10.1260/2047-4970.2.1.75

[29] Yu Gu and Jiangan Li. 2021. A novel WiFi gesture recognition method based on CNN-LSTM and channel attention. In *Proceedings of the 3rd International Conference on Advanced Information Science and System (AISS'21)*. ACM, Article 85, 4 pages. https://doi.org/10.1145/3503047.3503148

[30] Gido Hakvoort. 2013. The immersive museum. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces (ITS'13)*. ACM, 463–468. https://doi.org/10.1145/2512349.2514598

[31] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool release: Gathering 802.11n traces with channel state information. *SIGCOMM Comput. Commun. Rev.* 41, 1 (Jan. 2011), 53. https://doi.org/10.1145/1925861.1925870

[32] Ying He, Yan Chen, Yang Hu, and Bing Zeng. 2020. WiFi Vision: Sensing, Recognition, and Detection With Commodity MIMO-OFDM WiFi. *IEEE Internet Things J.* 7, 9 (Sept. 2020), 8296–8317. https://doi.org/10.1109/JIOT.2020.2989426

[33] Atsushi Hiyama, Jun Yamashita, Hideaki Kuzuoka, Koichi Hirota, and Michitaka Hirose. 2004. Position tracking using infra-red signals for museum guiding system. In *Proceedings of the 2nd International Conference on Ubiquitous Computing Systems (UCS'04)*, Hitomi Murakami, Hideyuki Nakashima, Hideyuki Tokuda, and Michiaki Yasumura (Eds.). Springer-Verlag, Berlin, 49–61. https://doi.org/10.1007/11526858_5

[34] Anna Huang, Dong Wang, Run Zhao, and Qian Zhang. 2019. Au-Id: Automatic user identification and authentication through the motions captured from sequential human activities using RFID. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 48 (jun 2019), 26 pages. https://doi.org/10.1145/3328919

[35] Çağrı Imamoğlu and Aslı Canan Yılmazsoy. 2009. Gender and locality-related differences in circulation behavior in a museum setting. *Museum Manage. Curator.* 24, 2 (June 2009), 123–138. https://doi.org/10.1080/09647770902857539

[36] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom'18)*. ACM, 289–304. https://doi.org/10.1145/3241539.3241548

[37] Robert J. Johnston. 1998. Exogenous Factors and Visitor Behavior: A Regression Analysis of Exhibit Viewing Time. *Environ. Behav.* 30, 3 (May 1998), 322–347. https://doi.org/10.1177/001391659803000304

[38] Joel Lanir, Tsvi Kuflik, Julia Sheidin, Nisan Yavin, Kate Leiderman, and Michael Segal. 2017. Visualizing museum visitors' behavior: Where do they go and what do they do there? *Pers. Ubiquit. Comput.* 21, 2 (Apr. 2017), 313–326. https://doi.org/10.1007/s00779-016-0994-9

[39] Mingyue Li, Yougen Xu, and Zhiwen Liu. 2021. Direction of Arrival Estimation Using One-dimensional Convolutional Neural Network and Gated Recurrent Unit. In *Proceedings of the 3rd International Symposium on Signal Processing Systems (SSPS'21)*. ACM, 38–43. https://doi.org/10.1145/3481113.3481116

[40] Xiukui Li. 2019. A GPS-Based Indoor Positioning System With Delayed Repeaters. *IEEE Trans. Vehic. Technol.* 68, 2 (Feb. 2019), 1688–1701. https://doi.org/10.1109/TVT.2018.2889928

[41] Xian Li. 2020. Space-time distribution model of visitor flow in tourism culture construction via back propagation neural network model. *Pers. Ubiquit. Comput.* 24, 2 (Apr. 2020), 223–235. https://doi.org/10.1007/s00779-019-01342-w

[42] Wei-Chuan Lin, Winston K. G. Seah, and Wei li. 2011. Exploiting radio irregularity in the Internet of Things for automated people counting. In *Proceedings of the IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 1015–1019. https://doi.org/10.1109/PIMRC.2011.6139649

[43] Shangqing Liu, Yanchao Zhao, Fanggang Xue, Bing Chen, and Xiang Chen. 2019. DeepCount: Crowd Counting with WiFi via Deep Learning. (Mar. 2019). https://doi.org/10.48550/arXiv.1903.05316

[44] Yongsen Ma, Gang Zhou, and Shuangquan Wang. 2020. WiFi sensing with channel state information: A survey. *ACM Comput. Surv.* 52, 3, Article 46 (May 2020), 36 pages. https://doi.org/10.1145/3310194

[45] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign language recognition using WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 23 (Mar. 2018), 21 pages. https://doi.org/10.1145/3191755

[46] Payal T. Mahida, Seyed Shahrestani, and Hon Cheung. 2017. Localization techniques in indoor navigation system for visually impaired people. In *Proceedings of the 17th International Symposium on Communications and Information Technologies (ISCIT'17)*. IEEE, 1–6. https://doi.org/10.1109/ISCIT.2017.8261229

[47] Mark T. Marshall. 2018. Interacting with Heritage: On the Use and Potential of IoT Within the Cultural Heritage Sector. In *Proceedings of the 5th International Conference on Internet of Things: Systems, Management and Security*. IEEE, 15–22. https://doi.org/10.1109/IoTSMS.2018.8554899

[48] Johannes Meyer, Adrian Frank, Thomas Schlebusch, and Enkeljeda Kasneci. 2021. A CNN-based Human Activity Recognition System Combining a Laser Feedback Interferometry Eye Movement Sensor and an IMU for Context-aware Smart Glasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 172 (Dec. 2021), 24 pages. https://doi.org/10.1145/3494998

[49] Vincenzo Mighali, Giuseppe Del Fiore, Luigi Patrono, Luca Mainetti, Stefano Alletto, Giuseppe Serra, and Rita Cucchiara. 2015. Innovative IoT-aware services for a smart museum. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*. ACM, 547–50. https://doi.org/10.1145/2740908.2744711

[50] Theano Moussouri and George Roussos. 2015. Conducting Visitor Studies Using Smartphone-Based Location Sensing. *J. Comput. Cult. Herit.* 8, 3, Article 12 (May 2015), 16 pages. https://doi.org/10.1145/2677083

[51] Kai Niu, Fusang Zhang, Dan Wu, and Daqing Zhang. 2021. Exploring stability in WiFi sensing system based on fresnel zone model. *J. Front. Comput. Sci. Technol.* 15, 1 (2021), 60–72. https://doi.org/10.3778/j.issn.1673-9418.1912017

[52] Gretchen Nurse Rainbolt, Jacob A. Benfield, and Ross J. Loomis. 2012. Visitor Self-Report Behavior Mapping as a Tool for Recording Exhibition Circulation. *Visitor Studies* 15, 2 (Oct. 2012), 203–216. https://doi.org/10.1080/10645578.2012.715035

[53] Maxime Portaz, Matthias Kohl, Georges Quénot, and Jean-Pierre Chevallet. 2017. Fully Convolutional Network and Region Proposal for Instance Identification with Egocentric Vision. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW'17)*. IEEE, 2383–2391. https://doi.org/10.1109/ICCVW.2017.281

[54] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2.0: Passive human tracking with a single WiFi link. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'18)*. ACM, 350–361. https://doi.org/10.1145/3210240.3210314

[55] Kun Qian, Chenshu Wu, Zimu Zhou, Yue Zheng, Zheng Yang, and Yunhao Liu. 2017. Inferring Motion Direction using Commodity Wi-Fi for Interactive Exergames. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'17)*. ACM, 1961–1972. https://doi.org/10.1145/3025453.3025678

[56] Francesco Ragusa, Antonino Furnari, Sebastiano Battiato, Giovanni Signorello, and Giovanni Maria Farinella. 2019. Egocentric Visitors Localization in Cultural Sites. *J. Comput. Cult. Herit.* 12, 2, Article 11 (June 2019), 19 pages. https://doi.org/10.1145/3276772

[57] Aditya Singh Rathore, Weijin Zhu, Afee Daiyan, Chenhan Xu, Kun Wang, Feng Lin, Kui Ren, and Wenyao Xu. 2020. SonicPrint: A generally adoptable and secure fingerprint biometrics in smart devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services (MobiSys'20)*. ACM, 121–134. https://doi.org/10.1145/3386901.3388939

[58] Wilson Sakpere, Oshin Michael Adeyeye, and Nhlanhla B. W. Mlitwa. 2017. A state-of-the-art survey of indoor positioning and navigation systems and technologies. *South Afr. Comput. J.* 29, 3 (Dec. 2017), 145–197. https://doi.org/10.18489/sacj.v29i3.452

[59] Alexandra M. Schautz, Esther M. van Dijk, and Anke Meisert. 2016. The Use of Audio Guides to Collect Individualized Timing and Tracking Data in a Science Center Exhibition. *Visitor Studies* 19, 1 (Apr. 2016), 96–116. https://doi.org/10.1080/10645578.2016.1144032

[60] Daniel Schmitt and Michel Labour. 2021. Making sense of visitors' sense-making experiences: The REMIND method. *Museum Manage. Curator.* (Mar. 2021), 1–17. https://doi.org/10.1080/09647775.2021.1897953

[61] Beverly Serrell. 1997. Paying Attention: The Duration and Allocation of Visitors' Time in Museum Exhibitions. *Curator: Museum J.* 40, 2 (June 1997), 108–125. https://doi.org/10.1111/j.2151-6952.1997.tb01292.x

[62] Mamoona Birkhez Shami, Salman Maqbool, Hasan Sajid, Yasar Ayaz, and Sen-Ching Samson Cheung. 2019. People Counting in Dense Crowd Images Using Sparse Head Detections. *IEEE Trans. Circ. Syst. Video Technol.* 29, 9 (Sept. 2019), 2627–2636. https://doi.org/10.1109/TCSVT.2018.2803115

[63] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. 2021. WiFi-Enabled User Authentication through Deep Learning in Daily Activities. *ACM Trans. Internet Things* 2, 2, Article 13 (May 2021), 25 pages. https://doi.org/10.1145/3448738

[64] Jeffrey K. Smith and Lisa F. Smith. 2001. Spending Time on Art. *Empir. Studies Arts* 19, 2 (July 2001), 229–236. https://doi.org/10.2190/5MQM-59JH-X21R-JN5J

[65] Lisa F. Smith, Jeffrey K. Smith, and Pablo P. L. Tinio. 2017. Time spent viewing art and reading labels. *Psychol. Aesthet. Creat. Arts* 11, 1 (2017), 77–85. https://doi.org/10.1037/aca0000049

[66] K. Sornalatha and V. R. Kavitha. 2017. IoT based smart museum using Bluetooth Low Energy. In *Proceedings of the 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB'17)*. IEEE, 520–523. https://doi.org/10.1109/AEEICB.2017.7972368

[67] Gridaphat Sriharee, Niti Boonyakaite, Wachiraphan Charoeanwet, Pattarakorn Summat, and Pimphaka Saikhow. 2018. Integrating Bluetooth-Low Energy to Support Visitor Recommendation and Space Management. In *Proceedings of the VII International Conference on Network, Communication and Computing (ICNCC'18)*. ACM, 165–170. https://doi.org/10.1145/3301326.3301339

[68] Rolf Steier, Palmyre Pierroux, and Ingeborg Krange. 2015. Embodied interpretation: Gesture, social interaction, and meaning making in a national art museum. *Learn. Cult. Soc. Interact.* 7 (Dec. 2015), 28–42. https://doi.org/10.1016/j.lcsi.2015.05.002

[69] G. Styliaras, C. Constantinopoulos, P. Panteleris, D. Michel, N. Pantzou, K. Papavasileiou, K. Tzortzi, A. Argyros, and D. Kosmopoulos. 2020. The MuseLearn Platform: Personalized Content for Museum Visitors Assisted by Vision-Based Recognition and 3D Pose Estimation of Exhibits. In *Artificial Intelligence Applications and Innovations*, Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis (Eds.). Springer, Cham, 439–451. https://doi.org/10.1007/978-3-030-49161-1_37

[70] Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. 2011. Museum guide 2.0-an eye-tracking based personal assistant for museums and exhibits. In *Proceedings of the International Conference on Re-Thinking Technology in Museums*.

[71] Ayça Turgay Zıraman and Çağrı Imamoğlu. 2020. Visitor Attention in Exhibitions: The Impact of Exhibit Objects' Ordinal Position, Relative Size, and Proximity to Larger Objects. *Environ. Behav.* 52, 4 (May 2020), 343–370. https://doi.org/10.1177/0013916518804017

[72] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI'17)*. ACM, 216–220. https://doi.org/10.1145/3136755.3136817

[73] Raghav H. Venkatnarayan, Griffin Page, and Muhammad Shahzad. 2018. Multi-user gesture recognition using WiFi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'18)*. ACM, 401–413. https://doi.org/10.1145/3210240.3210335

[74] Edward Verbree, Sisi Zlatanova, Karl van Winden, Eva van der Laan, Antigoni Makri, Taizhou Li, and Haojun Ai. 2013. To localise or to be localised with WiFi in the Hubei museum? *Int. Arch. Photogram. Remote Sens. Spatial Info. Sci.* XL-4/W4 (Dec. 2013), 31–35. https://doi.org/10.5194/isprsarchives-XL-4-W4-31-2013

[75] F. Wahl, M. Milenkovic, and O. Amft. 2012. A Distributed PIR-based Approach for Estimating People Count in Office Environments. In *Proceedings of the IEEE 15th International Conference on Computational Science and Engineering*. IEEE, 640–647. https://doi.org/10.1109/ICCSE.2012.92

[76] Hao Wang, Daqing Zhang, Junyi Ma, Yasha Wang, Yuxiang Wang, Dan Wu, Tao Gu, and Bing Xie. 2016. Human respiration detection with commodity WiFi devices: Do user location and body orientation matter? In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16)*. ACM, 25–36. https://doi.org/10.1145/2971648.2971744

[77] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of WiFi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom'15)*. ACM, 65–76. https://doi.org/10.1145/2789168.2790093

[78] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom'14)*. ACM, 617–628. https://doi.org/10.1145/2639108.2639143

[79] Dan Wu, Daqing Zhang, Chenren Xu, Yasha Wang, and Hao Wang. 2016. WiDir: Walking direction estimation using wireless signals. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16)*. ACM, 351–362. https://doi.org/10.1145/2971648.2971658

[80] Wei Xi, Jizhong Zhao, Xiang-Yang Li, Kun Zhao, Shaojie Tang, Xue Liu, and Zhiping Jiang. 2014. Electronic frog eye: Counting crowd using WiFi. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'14)*. IEEE, 361–369. https://doi.org/10.1109/INFOCOM.2014.6847958

[81] Fu Xiao, Zhengxin Guo, Yingying Ni, Xiaohui Xie, Sabita Maharjan, and Yan Zhang. 2019. Artificial intelligence empowered mobile sensing for human flow detection. *IEEE Netw.* 33, 1 (Jan. 2019), 78–83. https://doi.org/10.1109/MNET.2018.1700356

[82] Steven S. Yalowitz and Kerry Bronnenkant. 2009. Timing and Tracking: Unlocking Visitor Behavior. *Visitor Studies* 12, 1 (Apr. 2009), 47–64. https://doi.org/10.1080/10645570902769134

[83] Yanni Yang, Jiannong Cao, Xuefeng Liu, and Xiulong Liu. 2018. Wi-count: Passing people counting with COTS WiFi devices. In *Proceedings of the 27th International Conference on Computer Communication and Networks (ICCCN'18)*. IEEE, 1–9. https://doi.org/10.1109/ICCCN.2018.8487420

[84] Yanni Yang, Jiannong Cao, Xiulong Liu, and Xuefeng Liu. 2020. Door-Monitor: Counting In-and-Out Visitors With COTS WiFi Devices. *IEEE Internet Things J.* 7, 3 (Mar. 2020), 1704–1717. https://doi.org/10.1109/JIOT.2019.2953713

[85] Yuji Yoshimura, Stanislav Sobolevsky, Carlo Ratti, Fabien Girardin, Juan Pablo Carrascal, Josep Blat, and Roberta Sinatra. 2014. An Analysis of Visitors' Behavior in the Louvre Museum: A Study Using Bluetooth Data. *Environ. Plan. B: Plan. Design* 41, 6 (Dec. 2014), 1113–1131. https://doi.org/10.1068/b130047p

[86] Yinggang Yu, Dong Wang, Run Zhao, and Qian Zhang. 2019. RFID based real-time recognition of ongoing gesture with adversarial learning. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems (SenSys'19)*. ACM, 298–310. https://doi.org/10.1145/3356250.3360045

[87] Xenophon Zabulis, Dimitris Grammenos, Thomas Sarmis, Konstantinos Tzevanidis, Pashalis Padeleris, Panagiotis Koutlemanis, and Antonis A. Argyros. 2013. Multicamera human detection and tracking supporting natural interaction with large-scale displays. *Mach. Vision Appl.* 24 (Feb. 2013), 319–336. https://doi.org/10.1007/s00138-012-0408-6

[88] Youwei Zeng, Dan Wu, Jie Xiong, Jinyi Liu, Zhaopeng Liu, and Daqing Zhang. 2020. MultiSense: Enabling multi-person respiration sensing with commodity WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 102 (Sep. 2020), 29 pages. https://doi.org/10.1145/3411816

[89] Fusang Zhang, Kai Niu, Jie Xiong, Beihong Jin, Tao Gu, Yuhang Jiang, and Daqing Zhang. 2019. Towards a Diffraction-based Sensing Approach on Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 33 (Mar. 2019), 25 pages. https://doi.org/10.1145/3314420

[90] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with WiFi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'19)*. ACM, 313–325. https://doi.org/10.1145/3307334.3326081

[91] Rui Zhou, Ziyuan Gong, Xiang Lu, and Yang Fu. 2020. WiFlowCount: Device-Free People Flow Counting by Exploiting Doppler Effect in Commodity WiFi. *IEEE Syst. J.* 14, 4 (Dec. 2020), 4919–4930. https://doi.org/10.1109/JSYST.2019.2961735

[92] Han Zou, Yuxun Zhou, Jianfei Yang, and Costas J. Spanos. 2018. Device-free occupancy detection and crowd counting in smart buildings with WiFi-enabled IoT. *Energy Build.* 174 (Sept. 2018), 309–322. https://doi.org/10.1016/j.enbuild.2018.06.040

Received 27 July 2021; revised 30 March 2022; accepted 4 April 2022