RFWash: A Weakly Supervised Tracking of Hand Hygiene Technique

Abdelwahed Khamis
UNSW, Sydney, Australia
a.khamiss@unsw.edu.au

Branislav Kusy
CSIRO, Brisbane, Australia
brano.kusy@data61.csiro.au

Chun Tung Chou
UNSW, Sydney, Australia
c.t.chou@unsw.edu.au

Mary-Louise McLaws
UNSW, Sydney, Australia
m.mclaws@unsw.edu.au

Wen Hu
UNSW, Sydney, Australia
wen.hu@unsw.edu.au

ABSTRACT

Each year, hundreds of thousands of people contract Healthcare Associated Infections (HAIs). Poor hand hygiene compliance among healthcare workers is thought to be the leading cause of HAIs and methods were developed to measure compliance. Surprisingly, human observation is still considered the gold standard for measuring compliance by World Health Organization (WHO). Moreover, no automated solutions exist for monitoring hand hygiene techniques, such as "how to hand rub" technique by WHO. In this paper, we introduce RFWash; the first radio-based device-free system for monitoring Hand Hygiene (HH) technique. On the technical level, HH gestures are performed back-to-back in a continuous sequence and pose a significant challenge to conventional two-stage gesture detection and recognition approaches. We propose a deep model that can be trained on unsegmented naturally-performed HH gesture sequences. RFWash evaluation demonstrates promising results for tracking HH gestures, achieving gesture error rate of $< 8\%$ when trained on 10-second segments, which reduces manual labelling overhead by $\approx 67\%$ compared to fully supervised approach. The work is a step towards practical RF sensing that can reliably operate inside future healthcare facilities.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**.

KEYWORDS

Hand Hygiene; Contactless Sensing; Millimeter Waves; Radar

ACM Reference Format:
Abdelwahed Khamis, Branislav Kusy, Chun Tung Chou, Mary-Louise McLaws, and Wen Hu. 2020. RFWash: A Weakly Supervised Tracking of Hand Hygiene Technique. In *The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3384419.3430733

SenSys '20, November 16–19, 2020, Virtual Event, Japan
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7590-0/20/11.
https://doi.org/10.1145/3384419.3430733

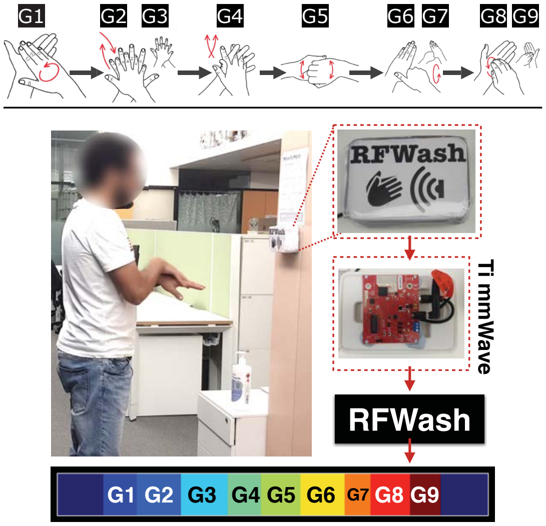


Figure 1: FRWash can monitor alcohol-based handrub procedure recommended by the WHO. The 9 steps are marked by the labels G1, G2 etc. Check system demo here https://youtu.be/t28NXk9XABE

1 INTRODUCTION

Healthcare Associated Infections (HAIs) find their way to one in twenty five patients admitted to hospitals [2] and continue to lead to increased patient mortality and healthcare cost [2]. Proper hand hygiene protocol, i.e. frequent and thorough hand cleaning, is an effective way to combat HAIs [8]. This leads to the question of how one can monitor hand hygiene (HH) adherence in an hospital environment. The conventional approach for HH adherence monitoring is to employ a team of observers (e.g., overt nurse trained auditors) to record Hand Hygiene Opportunities (HHOs) and the number of times health care workers (HCWs) comply with the protocol. Today, this is considered to be the gold standard for measuring compliance by the World Health Organization (WHO).

Attempts to implement automated alternatives for monitoring HH had a limited success so far. For example, electronic counters [23] and RFID [29] simply count hand washing activities. These tools provide a very limited picture of HH adherence. They cannot reveal whether hand hygiene technique — such as the nine-step

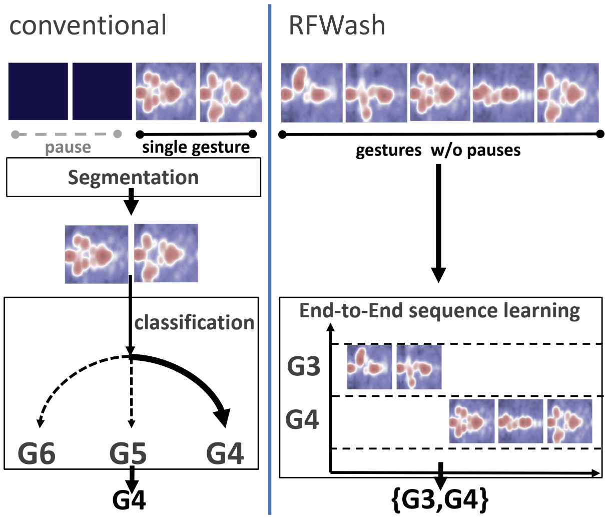


Figure 2: *Gesture sequence recognition:* Unlike conventional gesture recognition (left), the proposed model (right) is trained on unsegmented hand hygiene gestures and predicts labels for whole sequences of gestures in run-time.

procedure for applying alcohol-based handrub recommenced by WHO [1], see Fig. 1 (top row)— has been thoroughly adhered to. Although there are commercial camera systems for training HCWs to learn the correct HH technique, to the best of our knowledge, there exists no solution for automated monitoring of the HH technique in healthcare facilities.

This paper proposes to utilize commercial-off-the-shelf mmWave sensors to monitor the HH technique as shown in Fig. 1. Our vision is to embed these sensors at the alcohol-based handrub dispensers, which are distributed throughout the hospitals, to monitor whether HCWs have adhered to HH technique. Our vision will therefore enable much more fine-grained monitoring of HH adherence. The HH technique in Fig. 1 can be decomposed into 9 different hand movement patterns. While major progress has been made in gesture recognition using radio frequency (RF) signals recently [22], HH gesture monitoring presents unique challenges. First, six out of the nine steps of hand rub are very similar in that they are comprised of motions with the left and right hands mirrored. Second, some gestures are performed with two hands interlocked. Finally, the entire procedure is performed without a pause between consecutive gestures. Contiguous sequences of gestures have not been investigated in RF sensing literature before. In fact, previous RF-based sensing approaches [22] rely on pauses between gestures, which are employed as physical markers identifying the start and end of each motion segment. This approach trivially achieves accurate segmentation and the problem reduces to gesture classification. *Without enforcing the pauses, joint segmentation and classification becomes a challenging task.*

Back-to-back gestures with no pauses defy traditional segmentation techniques. Due to the significant interdependence between segmentation and subsequent recognition, poor segmentation, as we will see in Sec. 3.1 , deteriorates classification performance. Hence the approach can't be adapted to hand hygiene tracking. While the challenge of RF-based contiguous gestures recognition has been recognized in prior work [22, 32], to the best of our knowledge, no attempts were made to address it. For example, a WiFi-based sign language recognition system SignFi [22] addresses the segmentation issue by making the assumption that "manually segmented" single-gesture samples can be acquired. The assumption is unrealistic and contiguous gesture recognition "introduces many challenges"[22] to the previous approaches.

In this work, we address the problem by introducing RFWash; a segmentation-free approach for recognizing back-to-back HH gestures sequence. We draw inspiration from modern end-to-end speech recognition systems, which are similar to our problem because it is difficult to label continuous speech data. Of particular relevance to our problem are weakly supervised methods that can learn directly form data without requiring explicit data segmentation and full annotation. To this end, we develop a model that can be trained on back-to-back gesture sequences (Fig. 2) without requiring gesture segmentation, which can also reduce labelling overhead substantially.

A straightforward adaptation of sequence learning, however, doesn't work for long HH gestures sequences. Long training sequences pose two major challenges that RFWash needs to overcome. First, working with longer sequences leads to fewer training data points as a fixed-size training set gets split into fewer sub-sequences proportionally to the sub-sequence length. Second, the number of possibilities to align minimal gesture label sequence within an RF HH data sequence grows exponentially with sequence length [19]. Ultimately, the situation becomes ill-posed and results in poor alignment(Sec. 5.2). To tackle this, we use data augmentation to significantly increase the number of training samples without modifying the sequence content. Consequently, a significant improvement of sequence learning is realized. This paper makes the following contributions:

(1) We propose and implement **RFWash**, which is the first RF-based system for device-free monitoring of the nine-step Alcohol-Based Hand Rub (ABHR) technique.
(2) We **characterize the challenges of recognizing back-to-back HH gestures** using an RF-based gesture recognition processing pipeline. In particular, the lack of pauses between gestures makes segmentation difficult which, in turn, affects the performance of the subsequent classification component.
(3) We propose a **new sequence learning approach** that performs segmentation and recognition simultaneously. The model can be trained using continuous stream of minimally labelled RF data corresponding to naturally performed hand-rub gestures. We further extend the approach using **a novel data augmentation technique** to enable training on longer segments that are less labour intensive.
(4) We **extensively evaluate the performance of RFWash** using a dataset of 1,800 gesture samples collected from ten subjects over 3 months.

2 MOTIVATION

2.1 HH Monitoring in Real World

An ideal automated system for monitoring HH compliance should be able to detect attempts by HCW to perform hand rub procedures to track HH opportunities and to establish compliance rate baselines.

Table 1: Overview of automated HH monitoring systems.

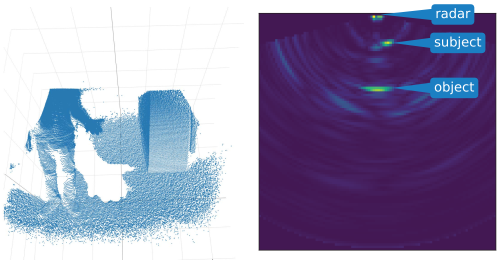
Work	Contact-Free	Hygiene Tech.	Inside Wards
Electronic Counters [23]	✓	✗	✓
RFID [29]	✗	✗	✓
Wearable [16]	✗	✓	✗(pathogens)
RGB Camera [20]	✓	✓	✗(privacy)
Depth Camera [12]	✓	✗	✗(privacy)
Depth Camera [44]	✓	✓	✗(privacy)
RFWash (radar-on-chip)	✓	✓	✓

Additionally, the system should monitor fine-grained parameters of HH technique (Fig. 1) itself. Such information can provide useful insights and help establish compliance rates of the healthcare facilities. The system must be capable of running unattended in real-world healthcare facilities. Prior study of 789 clinicians in a 380-bed tertiary hospital [14] has shown that automated HH training systems have limited impact on HH compliance as they do not operate inside wards. The benefits of an automated monitoring system that evaluates HH in-situ are therefore twofold. It will lead to improved compliance by reducing the Hawthorne effect [34] (i.e. inflation in hand hygiene compliance rates caused by behaviour change due to awareness of being observed by auditors) and it will provide quantitative data about hygiene quality within the healthcare facility. Despite the advent of machine learning algorithms for vision-based systems, the golden standard for assessing the HH in clinical facilities is direct human observation, which can only monitor a small fraction of hand hygiene opportunities [29]. A complete and automated HH monitoring system is yet to be realized. Table 1 surveys the key characteristics of current research-based and commercial automated solutions. All existing solutions perform well only on one or two aspects. More importantly, no solutions exist for monitoring the hand washing/rubbing process (i.e. 9-step HH technique recommended by the WHO) inside wards.

2.2 RF Sensing for HH monitoring

Practical solutions for monitoring HH opportunities inside hospitals wards include electronic counters [23] and RFID [29]. Counters simply count the washing dispenser activities to infer compliance rates, while RFID systems [29] track hygiene events by the proximity of healthcare workers from washing dispensers using wearable RFID tags. RFID is a proximity-based solution with a typical 1-meter location error and can miss more than 80% of hygiene events. To provide better localization accuracy, a network of cameras for tracking staff inside hospital was proposed [12]. However, neither approach can track the actual hand rub technique.

While no commercial solutions currently exist for tracking hand rubbing inside hospital wards, research solutions based on camera technology have been proposed [20, 44]. Since privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) limit the use of cameras in healthcare settings [12], camera-based systems employ image anonymization techniques. One such example [12] uses a depth camera that conceals color information as each pixel value in depth image represents the distance between the pixel and the camera instead of the color. Using depth camera alone does not provide sufficient privacy guarantees. Despite careful control of the field of view of the cameras and reduced image resolution [7], the images can still show detailed visual appearance of a person that may be used to track her and invade her privacy.



(a) Depth camera (b) RF ranging heatmap

Figure 3: Genuinely anonymous signal: RF Ranging data has much lower resolution than frames from co-located Kinect.

Fig. 3 compares the RF signal from a TI mmWave radar used in this paper to the depth data from a co-located Kinect depth camera. The camera was mounted in a way that prevents capturing the subject's face. While both devices provide ranging information (i.e., how far objects from the sensing device), the RF heatmap has a significantly lower personal information content, significantly reducing the risk of privacy intrusion. We believe that the value of RF sensing can contribute to many other privacy-sensitive healthcare applications such as ICU activity logging [7, 18].Compared to other RF sensing technologies like WiFi and RFID, the mmWave is **self-contained** (i.e. sender/receiver in one on-the-chip device), with a **small form factor** can be easily housed in a hand sanitizer bracket, **less vulnerable to interference** [28] and **highly sensitive to small motions** [13]. Additionally, the **better spatial resolution** of the mmwave can be leveraged to filter out irrelevant motions that are often present in the real-world due to other people or equipment. Together with privacy-protection property discussed above, these advantages make the mmWave radar an ideal candidate for large-scale adoption in real-life healthcare facilities.

3 TECHNICAL MOTIVATION

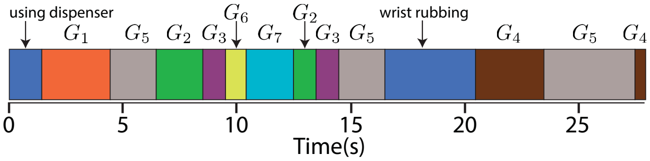


Figure 4: Timeline of HH technique of a practicing HCW.

HCWs are expected to execute the HH protocol at appropriate occasions at work, e.g. , before and after touching a patient. The hospitals facilitate this by placing soap or alcohol-based handrub dispensers at many easily accessible places in and outside of the wards. HCWs are expected to follow a standard hand cleaning procedure (Fig. 1) between 20-30 seconds to ensure their hands are thoroughly cleaned. To understand the current state of HH practices in healthcare environments, we conducted face-to-face interviews with active HCWs from Prince of Wales Hospital in Sydney. During the interview, we asked the HCWs to show us how they would typically execute their handrub procedure. We used a camera to record the process [1] and analyzed the video to obtain the handrub gesture sequence and timing information, see Fig. 4. The figure shows that the real-life gesture sequence diverges from the ideal expected sequence shown in Fig. 1 where gestures G_1, G_2, ..., G_9 are executed consecutively. Instead, Fig. 4 shows that gestures are repeated and are not in the expected order. This is because the HCW continued to rub her hands until all alcohol dried off her hands. Furthermore, we also note the timing variation for each gesture. This simple example illustrates the intricacies involved in hand rubbing. We expect significant deviation of the real-world hand rubbing from the ideal protocol.

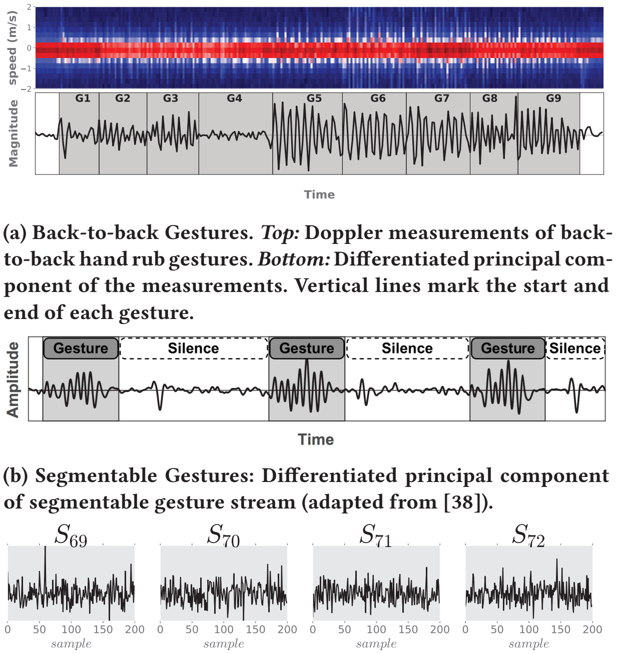
Based on the above observations, the first goal of RFWash is to accurately track the sequence of gesture poses performed by a HCW. The recorded sequence can be compared to the expected set of poses $G_1, G_2, \cdots, G_9$, e.g. by comparing the union of the detected poses to the set of expected poses. Additionally, RFWash tracks the timing information to help in assessing compliance against the 20-30 second duration guidelines. More complex compliance analysis based on the pose sequence and timing information could be done, but is beyond the scope of this paper.

3.1 Back-to-back Gesture Tracking

In this section, we explore the limitations of existing RF gesture processing algorithms in the HH scenario. Popular RF gesture recognition approaches follow a two-stage architecture, with detection/segmentation step followed by recognition step [4, 22, 37, 38]. Here, a critical assumption is that we can segment the RF time-series into segments where each segment contains one gesture only. Hence, a classifier can be trained and tested on these well-separated segments.

Typically, segmentation is done in one of two ways:

- *Gestures are naturally segmentable.* The users introduce a brief pause before and after performing each gesture [38], which makes the detection of the start and end of individual gestures simpler (Fig. 5b). Training samples either contain only relevant gesture data [32, 38] or the gesture data with additional samples that represent "no gesture" [27]. In the run-time, a segmentation module automatically segments gestures utilizing no motion or "silent" periods.
- *Users annotate continuous gestures manually.* Applications such as sign language recognition do not have segmentable gestures and the automated segmentation step from the previous approach fails. The limitation can be overcome by manual segmentation [22], i.e., manual extraction of segments, each of which contains a gesture. The key drawback of such approach is the high intensity of labour that the manual segmentation requires. The labelling of RF signals is not intuitive and can introduce errors compared to more natural modalities, such as video .



(a) Back-to-back Gestures. *Top:* Doppler measurements of back-to-back hand rub gestures. *Bottom:* Differentiated principal component of the measurements. Vertical lines mark the start and end of each gesture.

(b) Segmentable Gestures: Differentiated principal component of segmentable gesture stream (adapted from [38]).

(c) Manually Segmented Gestures: PCA of manually segmented sign language gestures samples from public dataset [22]

Figure 5: *Back-to-back* vs *segmentable* vs *manually segmented* gestures

Why do we propose to use a segmentation-free approach? Fig. 5a shows the Doppler measurements (top graph) and its differentiated principal component (bottom graph) of a real-life execution of the HH technique. The vertical lines in the bottom graph show the correct gesture boundaries. It shows that gesture boundaries are sharp with minimal period of "no gesture" samples in between. Therefore, threshold-based segmentation [32] fails to recognize gesture boundaries. Consequently, most segmented sequences contain RF signatures from multiple gestures. A classifier trained on such data will perform poorly.

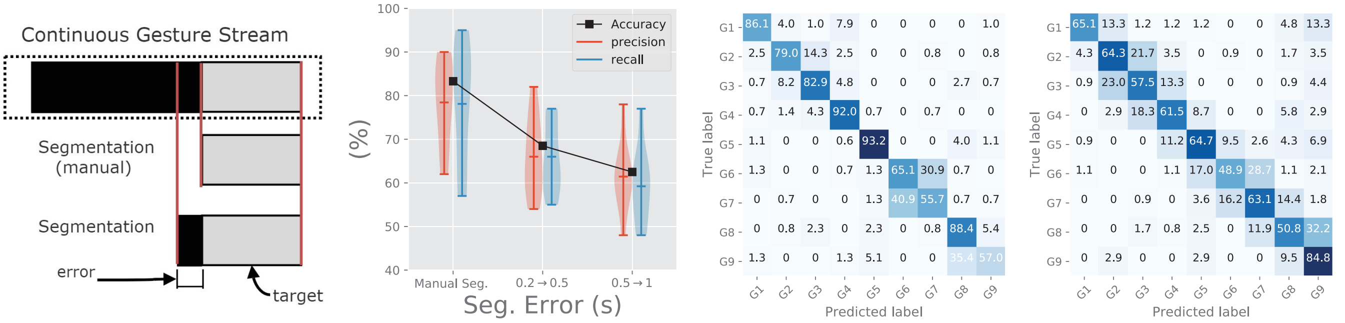
The impact of segmentation errors.

To quantify the errors due to inaccurate segmentation, we applied SignFi algorithm [22] to RF traces of HH gestures. The algorithm uses a deep CNN architecture originally designed to classify 276 sign language gestures, which we adapted to better suit our application scenario[2]. We evaluated SignFi on our dataset of naturally-performed HH technique from ten subjects using manually segmented samples that contain exactly one gesture in each segment[3]. Using two-second Doppler Range measurements and session-based cross-validation, we obtained accuracy of 83.3% (see Sec. 4.1 for the details). The confusion matrix (Fig. 6c) shows that the accuracy is more than 79% for most gestures except for some of the mirrored gestures, i.e., G_6/G_7 and G_9. Anecdotally, RF signatures of (G_6, G_7) and (G_8, G_9) are similar to each other and are more likely to result in incorrect classification.

[1] Ethical approval has been granted by the University of New South Wales (Approval Number HC180818)

[2] The convolutional layer in [22] has three $3x3$ kernels. This produced poor results on our Range Doppler measurements, hence, we increased the number of kernels from 3 to 512, which improved its performance significantly.

[3] Sample-level labelling was done using a synchronized camera.



(a) Segmentation error **(b) Imapct of of seg. error on Accuracy** **(c) cf matrix, manual seg.** **(d) cf matrix,seg. error:** $0.5 \rightarrow 1$s

Figure 6: Classification is highly dependent on segmentation quality in RF gesture recognition systems

Table 2: Time cost for manual and sequence labelling

Method	RF data	Labelling	Saving
Manual Segmentation	4 mins @ 8Hz	18 mins	-
10*s* Sequence labelling	4 mins @ 8Hz	6 mins	66.6%

To investigate the effect of segmentation error, we deliberately allowed segments to contain a few samples from neighbouring gestures. We ensured that the majority of the samples in a segment corresponded to the target gesture (see Fig. 6a for an illustration). In particular, we allow for an overlap of 1-25% and 25-50%, corresponding to 0.2-0.5 second and 0.5-1 second overlaps, respectively. This allowed us to study the impact of different levels of segmentation error on the classification accuracy. Fig. 6b shows that the accuracy gets worse when the segmentation error increases. This shows that SignFi does not handle the segmentation errors well.

The cost of manual segmentation and labelling. Since RF samples are difficult to label and segment directly, we used a synchronized video camera in our experiment. The gestures were identified in the video and labels were propagated to the corresponding RF signatures. Despite using the camera feed as a visual aid, we found the process very time-consuming and we investigated an alternative method for annotating RF segments.

Sequence labelling. We introduce a new approach, which we call *sequence labelling*, to reduce the complexity of manual labelling. Two key ideas in sequence labelling are: 1) We ask users to annotate relatively long continuous sequences of data; and 2) We request users to annotate gesture sequences without capturing the exact timing information of individual gesture boundaries.

Let us consider an example. Assume that we have a collection of 20 data frames $\{f_0, \ldots, f_{19}\}$ which contains the gestures $G1$, $G2$, and $G3$ in that order. Manual segmentation requires us to identify gesture boundaries or map each frame to a gesture, e.g. the annotated sequence is $G_1 \in [f_0, f_5], G_2 \in [f_6, f_{13}], G_3 \in [f_{14}, f_{19}]$. In contrast, sequence labelling will annotate this collection of frames simply as $G_1 \rightarrow G_2 \rightarrow G_3$, which says the order of gestures in the frames are G_1, G_2 and G_3 without having to specify transition times. The work to obtain sequence labelling is therefore lower.

We quantified the time required to perform manual segmentation and sequence labelling experimentally. We asked 3 annotators to label four minutes of RF data sampled at 8 Hz using these methods. The average time taken is shown in Table 2. On average, manual segmentation took 18 minutes while sequence labelling took only 6 minutes, resulting in a saving of $\approx 66.6\%$. We note that manual labelling and segmentation cost can be significantly higher for higher RF sampling rates such as 200Hz in [22] or 1kHz in [32, 37]. We will show in the next section that it is possible to achieve highly accurate gesture segmentation and classification, based on sequence labels. In contrast to classical supervised learning, which requires fully annotated data, our weakly supervised method only requires minimally labelled data.

Summary: The assumption of easily segmentable input that is commonly used by existing RF-based gesture recognition approaches does not hold in the HH gesture recognition scenario. We show that the HH gesture classification accuracy depends heavily on the segmentation quality. While good quality classifiers can be developed using manually segmented data, this incurs substantial labelling costs. Inspired by methods from speech and handwriting recognition literature, RFWash departs from the existing RF sensing segmentation approaches and proposes new methods to learn from weakly labelled *unsegmented data.*

4 RFWASH

Fig. 7 shows the architecture of the proposed RFWash framework. RFWash is trained on sequences of HH gestures in the RF space and their corresponding sequence labels. As discussed in the previous section, a sequence label only contains the order of the gestures in the segment. The training process, therefore, needs to determine the most likely mapping of gesture labels to each RF frame. This is done through the process we call *alignment learning.* At runtime, RFWash model internally assigns likelihood to each *(input RF frame, gesture)* pair, which is then used to infer the most likely gesture sequence. Before we delve into details about the model itself, we explain the input RF measurements in the next section.

4.1 RF Measurements

RFWash uses mmWave radar mounted on a soap dispenser, to collect RF signatures of subjects that perform hand cleaning. Fig. 8a shows the system setup. While many subjects may be present in a hospital environment, a subject that performs hand cleaning will stand close to the radar (e.g., within 1 m). The subject will face the

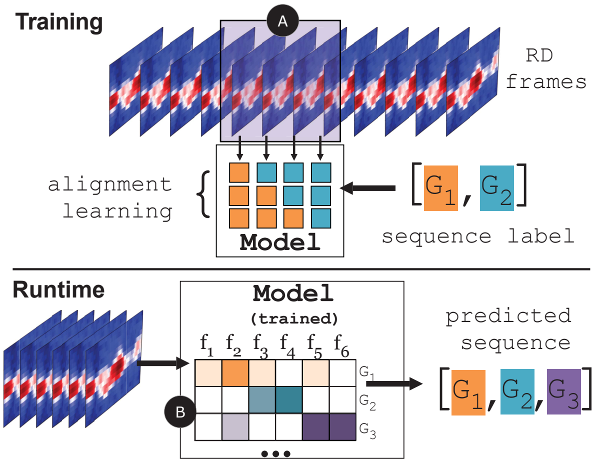


Figure 7: RFWash is trained on continuous RF samples (A) of HH gestures and corresponding sequence labels. The model automatically learns which frames correspond to individual gestures (e.g. G_1 vs G_2) through "alignment learning". In runtime, per-frame gesture predictions (B) are produced and used to estimate the most likely gesture sequence .

radar and her hands will be at approximately the same height as the radar. Consequently, our goal is to measure the velocity of her hand motions and filter out other irrelevant signals.

A mmWave radar transmits a sinusoidal wave $T(t)$, called a "chirp", of linearly changing frequency and a time delayed version of the transmitted signal is received for every reflector in the environment, including the hands of the subject washing her hands. Formally, the frequency of a chirp at time t can be expressed as:

$$f_t = f_0 + \frac{B}{T}t, \tag{1}$$

where f_0 is the starting frequency of the chirp, B is the bandwidth and T is the chirp duration. Let $A(t)$ be the amplitude of $T(t)$ at time t. The transmitted signal $T(t)$ can be expressed as :

$$T(t) = A(t)\sin(2\pi(f_0 t + \frac{B}{2T}t^2)). \tag{2}$$

When the transmitted signal is reflected by a stationary object at distance D_0 from the radar, the reflected signal $R(t)$ is:

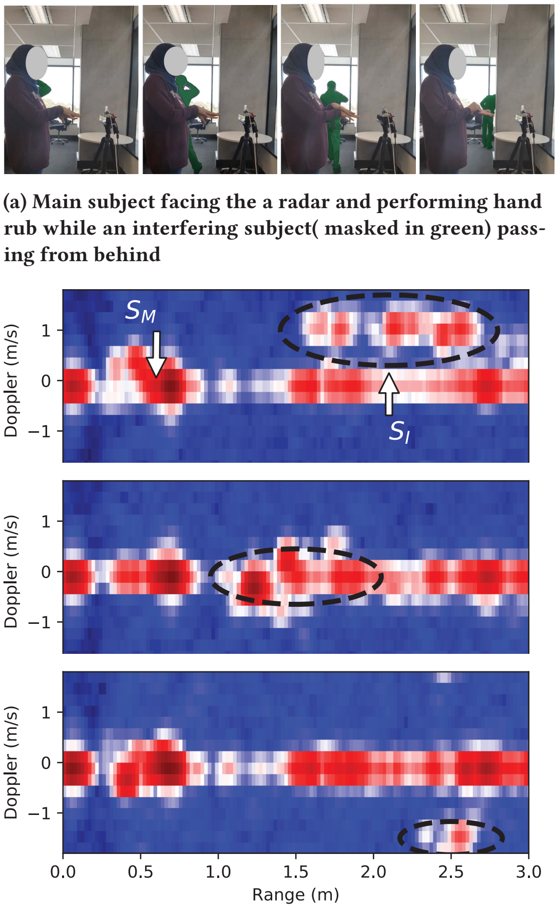
$$R(t) = E(t)\sin(2\pi(f_0(t - t_d) + \frac{B}{2T}(t - t_d)^2)), \tag{3}$$

where $E(t)$ is the amplitude modulated by the object, the round-trip time delay is $t_d = (2D_0)/c$ with c being the speed of light.

The signals $T(t)$ and $R(t)$ are mixed on the radar to produce the received signal $S(t)$. It can be shown that $S(t)$ has two frequency components: 1) the difference in frequencies between $T(t)$ and $R(t)$, 2) the sum of their frequencies. A low pass filter can be applied to remove the second component:

$$S(t) \approx C(t)\cos(2\pi(\frac{2BD_0}{cT}t + \frac{2f_0 D_0}{c})). \tag{4}$$

where $C(t)$ is the amplitude. The frequency of $S(t)$, which is given by $\frac{2BD_0}{cT}$, is called beat frequency and can be used to estimate the objects distance D_0.



(a) Main subject facing the a radar and performing hand rub while an interfering subject(masked in green) passing from behind

(b) Consecutive RD measurements frames showing main subject (S_M) can be separated from passing interfering subject (S_I) by range cutoff.

Figure 8: RD frames measurements

In general, there are multiple objects in the vicinity of the radar and the mixed received signal will contain multiple beat frequencies. We can resolve these with Fast Fourier Transform (FFT) and consequently compute distances between each object and the radar.

However, range alone does not provide sufficient information to solve our problem. The subject's hands during the handrub are very close to each other during the entire procedure. We need more information to differentiate the gestures. Fortunately, the mmWave radar allows us to measure Doppler frequency shift in the $S(t)$ signal, through which we can obtain the velocity of the objects moving in the scene.

We use mmWave signal $S(t)$ to derive *intensity map* of the scene shown in Fig. 8b. Intensity map $I(t, r, v)$ has the following interpretation: the intensity $I(t, r, v)$ is higher if there is a higher chance at time t of finding an object located at distance r from the radar and moving at speed v. Fig. 8b shows the intensity map at three different time instants, with r plotted from 0 to 3m, and v from -2 to 2 m/s. Large intensity is shown in red. We will refer to the intensity map $I(t, r, v)$ at a point in time as a Range-Doppler (RD) frame.

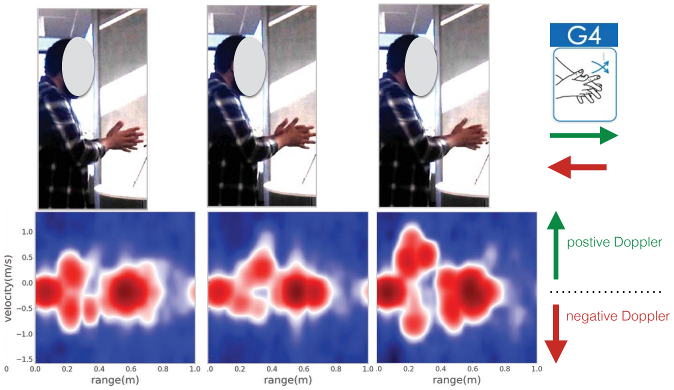


Figure 9: Example RD frames for gesture G_4. Video preview of the motion: https://youtu.be/t28NXk9XABE

RFWash needs to be robust to interference from nearby moving objects and people. Fig. 8a shows a subject performing handrub in front of the radar. The green person in the figure is within the range of the radar and acts an as interferer. Fig. 8b shows RD frames at three time instants, with the location of the subject's hands and interferer marked by S_M and S_I , respectively. We note that the intensity of all RD frames stay approximately unaffected in the S_M region, by the interferer's movement (dotted ellipses in Fig. 8b). From now on, we limit the range r to less than 1m so as to focus on the main subject only. For illustration, we show RD frames for one gesture (G_4) in which the user is rubbing palms with fingers interlaced. This motion pattern results in simultaneous Doppler change in the positive and negative directions as shown in Fig. 9. We perform interference removal and de-noising [40]. These steps remove the static reflection of the torso of the subject and amplify hand motions related to the gesture performed in the RD frame. The input to RFWash deep model is a stack of processed and normalized RD frames after applying cutoff at 1m to each frame and resizing them to 50×50 pixels.

4.2 Deep Learning Model

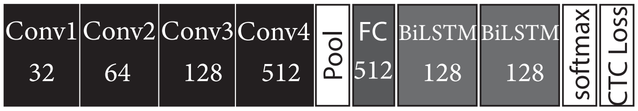


Figure 10: RFWash Network Architecture. All convolutions are 3×3 (the number of filters are denoted in each box).

The layering structure of our deep learning model is shown in Fig. 10. Convolution(Conv) layers followed by Max Pooling (2x2), Fully Connected (FC) and Bidirectional LSTM layers are employed for extracting spatiotemporal gesture features from input RD frames, a softmax layer. Connectionist Temporal Classification (CTC) is employed to predict the gesture sequence.

As illustrated in Fig. 7, RFWash takes a segment consisting of stack of consecutive T RD frames $X = [x_1, ..., x_T] \in \mathbb{R}^{50 \times 50 \times T}$ from continuous stream as input. The goal is to infer the gesture sequence ℓ performed by a HCW where $\ell = [\ell_1, \ell_2, .., \ell_K] \in \mathcal{A}^{1 \times K}$ where $\mathcal{A}$ is the set of possible gestures and $K \leq T$. Note that the continuous segment can contain irrelevant motions (i.e. user stationary or walking away from device) or gestures irrelevant to the nine poses (i.e. wrist washing). Thus we define an additional "no gesture" class G_{No} to handle irrelevant motions and other gestures. Ultimately, the set possible gestures $\mathcal{A} = \{G_{\text{No}}\} \cup \{G_1, ..., G_9\}$.

Recall that we use *sequence label* ℓ, rather than frame-by-frame label $\pi = [\pi_1, \cdots, \pi_T] \in \mathcal{A}^{1 \times T}$ to reduce the labelling cost (Sec. 3.1). π is also called *gesture path*. An associated challenge with ℓ is the lack of temporal alignment as it can be compatible with many plausible gesture paths. For example, if the sequence label is $G_1 \longrightarrow G_2$ for an input of 4 frames, the label is compatible with the gesture paths $[G_1, G_2, G_2, G_2]$, $[G_1, G_1, G_2, G_2]$ and $[G_1, G_1, G_1, G_2]$. Intuitively, the model resolves this challenge by considering the probability of all plausible gestures paths for a particular sequence label.

4.2.1 Spatiotemporal Feature Extraction. Motions captured by the mmWave radar in a single RD frame have identifiable spatial pattern on range and velocity dimensions. Additionally, the temporal dynamics of each gesture will be present in consecutive RD frames. We utilize spatiotemporal feature extraction layers composed of five Convolutional layers followed by a fully connected layer and two RNN (Recurrent Neural Network) layers. RNN achieve good performance in sequential data modeling and are a good choice for capturing temporal dynamics of the gestures. However, in the context of HH technique, the mirrored gestures discussed in Sec. 3.1, present unique challenges because of their similarity in RF domain. Therefore, we employ bidirectional recurrent layers with LSTM cell type (BiLSTM [31]) to enable the network to use all available input information in the past and the future from a specific RD frame. In this configuration, two separate recurrent layers running in the forward direction (future) and the backward direction (past) are utilized to learn the complex temporal dynamics.

The spatiotemporal feature extraction layers and softmax activation process input RD frames X to produce frame-wise probabilities of different gestures Y, which we call BiLSTM posterior. Y can be interpreted as the probability of observing A gestures across T frames. This is further processed by the temporal alignment to estimate the most likely gesture sequence.

4.2.2 Temporal Alignment Learning. RFWash implements alignment learning to infer the hand rub gesture sequence by mapping the output of BiLSTM components (i.e., BiLSTM posterior) to the corresponding gesture path. We rely on CTC algorithm [11], which is explained next in details.

Let $Y = [y_1, \cdots, y_T] \in \mathbb{R}^{A \times T}$ be the softmax-normalized BiLSTM output for a stack of T RD frames, where $A = |\mathcal{A} \cup \{\phi\}|$ where ϕ denotes a blank. A blank is used by CTC to account for the probability of observing 'no labels' and modeling the transition between gestures within sequence. This means $A = 11$ for RFWash. The vector $y_t, t \in \{1, \ldots, T\}$ can be interpreted as follows: $y_{t,k}$ denotes the probability that the gesture at time t is k where $k = 1, \ldots, A$.

We can calculate posterior probability for any gesture path $\pi = [\pi_1, \cdots, \pi_T]$ given the observations X as follows:

$$P(\pi|X) = \prod_{t=1}^{T} y_{t, \pi_t}, \forall \pi_t \in \mathcal{A}. \tag{5}$$

We note that the posterior probabilities obtained in Eq. (5) are conditionally independent for different gesture paths. This is desirable in our context, as we do not want the gesture classifier to be dependable on the order of gestures in the training data.

In the CTC framework, the probability of the sequence label ℓ is the sum of the probabilities of all its compatible gesture paths:

$$P(\ell|X) = \sum_{\{\pi|\mathcal{B}(\pi)=\ell\}} P(\pi|X), \tag{6}$$

where $\mathcal{B}$ is an operator that removes consecutive label repetitions and blanks in π. Intuitively, Eq. (6) considers all possible alignments in ℓ [4]. The most probable sequence label ℓ^* can be predicted as:

$$\ell^* = \operatorname*{argmax}_{\ell} P(\ell|X), \tag{7}$$

and the network can be trained using standard back-propagation method, minimizing the following:

$$\mathcal{L}(\ell, X) = -\log P(\ell|X). \tag{8}$$

For gesture timing estimation, we first process BiLSTM output to estimate the top gesture path $\hat{\pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_T]$ by selecting the gesture with the top probability at each frame . Then we set the starting time of a gesture to the frame with highest probability for that particular gesture. Finally, the end time is set to frame before the starting point of next gesture in sequence or the end of segment.

Data augmentation Up to this point, the model training can proceed using unsegmented input X of arbitrary lengths T and the corresponding sequence labels ℓ (Eq. (8)). Larger T will reduce the annotation effort as fewer sequences need to be annotated for a given training set. However, very long segments would result in a few training samples. Additionally, according to our experimental observation (Sec. 5.2), training on long sequences results in poor temporal alignment. Since we can't tamper with sub-sequences, RFWash employs "order preserving" concatenation of existing samples to augment training data that can lead to the increase of the number of samples quadratically. For example, let X_a and X_b be two stacks of RD frames, and their corresponding sequence labels are ℓ_a and ℓ_b. A new stack is obtained by concatenating X_a and X_b to form $[X_a, X_b]$ and its corresponding sequence label is $\mathcal{B}([\ell_a, \ell_b])$. Concatenation is applied on the training sequences from same and different users. We use the augmentation to increase dataset by the maximum of 10x. Prior to applying "order preserving" augmentation on the sequences level, we apply random jittering [35] on the the individual RD frames within each sequence. Jittering simulates the random noise that can be present in wireless radio environments and ensures no two sequences (after applying "order preserving augmentation") have the exact same copy of RD frame.

5 EVALUATION

We collected natural back-to-back handrub data where subjects move between gestures without pausing. This was followed in all data collection sessions. RFWash prototype uses TI mmWave IWR1443 sensor [3] that operates in the 60GHz to collect RF measurements at 8 Hz. RD measurements in our setup are exported with a ranging resolution of $\approx$ 4 cm and velocity resolution of $0.25ms^{-1}$. The deep model can process 10s segments (80 frames) within 57.5 ± 3.1 ms and 480 ± 20ms on, respectively, GPU (Nividia GeForce RTX 2080 Ti) and CPU (3.70 GHz Intel Xeon W-2135). Thus, the data can be processed at rate of 166 Hz on commodity hardware with a pure CPU implementation. We recruited ten subjects for data collection (8 males and 2 females)[5]. The subjects' heights ranged from 160cm to 193cm. We observe that the participant's height had a limited impact on the RD pattern captured in our experiments. A possible reason is the poor angular resolution of the radar in the vertical direction compared to horizontal direction (15° in azimuth and 58° in elevation)[3, 21] which causes motions towards/away from the radar to dominate. The subjects had no previous experience with the handrub procedure. They were shown a video of hand rubbing produced by the WHO, and they were asked to repeat the whole process two times before starting data collection.

In each session, a subject performed the handrub gestures from $G_1 \rightarrow G_9$ then stayed stationary or walked away from the device and returned. Note that subjects could miss one or more gesture. We collected a total of 1800 gesture samples (10 subjects × 4 sessions × 5 repetitions × 9 gestures). Also, we collected additional "unseen gesture sequence" dataset (Sec. 5.2.2) where the subjects perform handrub gestures in a random order. Time taken by subjects in each gesture is shown in Fig. 11a (average time is 3.2s). All sessions were recorded using synchronization between video and RF frames from radar was done using NTP. Later, a human auditor inspected the recorded video and labelled the video frame-by-frame. Following [39], we employ session-based cross validation for evaluation by default. In this scheme one hand rubbing session is held out for test and the rest are used for training.

Table 3: Gesture Error Rate (GER) examples

Ground truth	Prediction	GER	explanation
$[G_2]$	$[G_2]$	0%	-
$[G_2]$	$[G_1, G_2]$	100%	$\frac{\text{deletion}}{\text{length of ground truth}}$
$[G_2]$	$[G_1, G_3]$	200%	$\frac{\text{substitution+deletion}}{\text{length of ground truth}}$

5.0.1 Metrics. To evaluate the performance of RFWash, we use the following metrics:

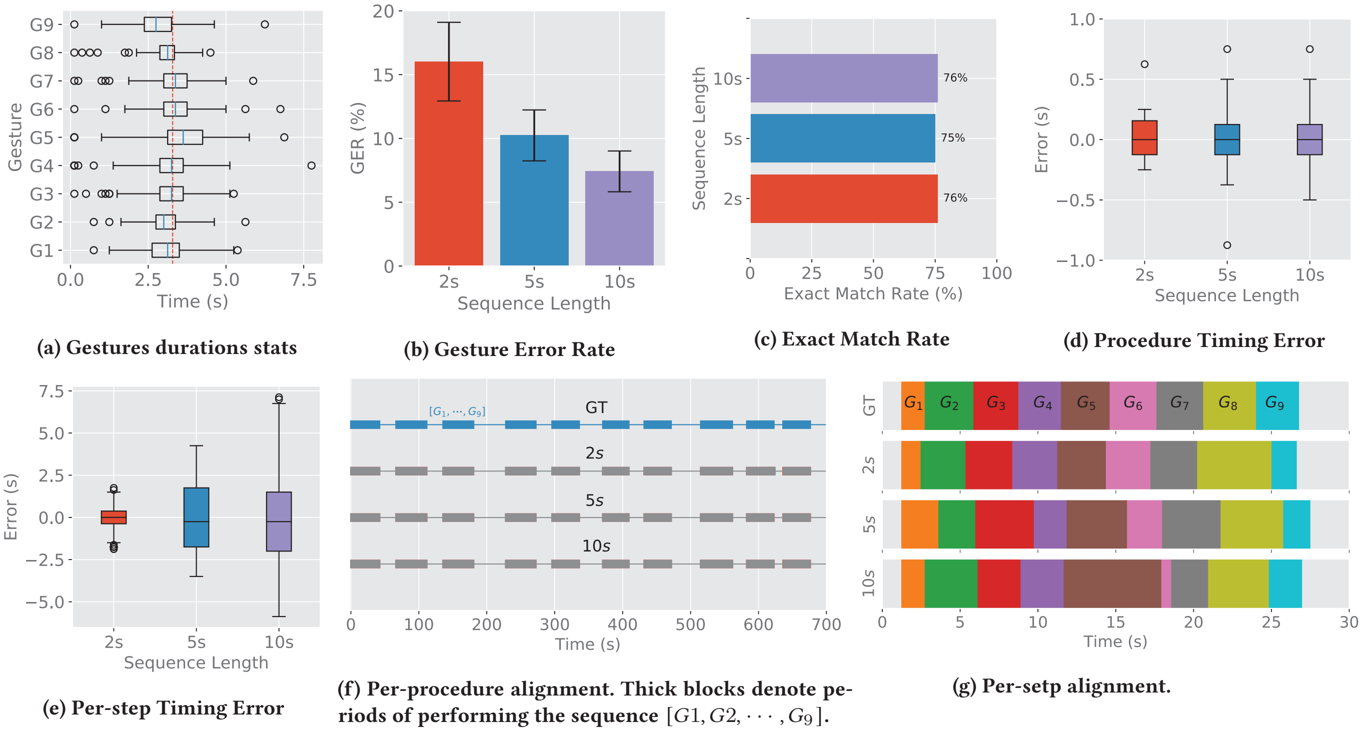
- **Gesture Error Rate (GER):** defined as the minimum number of gesture insertions, substitutions, and deletions needed to transform the predicted gesture sequence into the ground truth gesture sequence, divided by the number of gestures in the ground truth. Table 3 shows a few examples of GER. This metric mimics Word Error Rate (WER) which is a standard metric in sequence recognition problems.
- **Exact Match Rate (EMR):** defined as the percentage of predicted sequence that exactly match ground truth (i.e. sequences with GER = 0%).
- **Timing Error:** this is the difference between gesture estimated time and ground truth timing. Gesture timing error is calculated for gesture sequences with EMR of 100%.

5.1 Weakly Supervised Gesture Tracking

We evaluate the gesture sequence recognition of RFWash using the three metrics discussed earlier. Fig. 11b shows the mean GER when

[4] We note that the CTC forward backward algorithm is more efficient than considering all possibilities [11].

[5] Ethical approval has been granted by the University of New South Wales (Approval Number HC180818).



(a) Gestures durations stats

(b) Gesture Error Rate

(c) Exact Match Rate

(d) Procedure Timing Error

(e) Per-step Timing Error

(f) Per-procedure alignment. Thick blocks denote periods of performing the sequence $[G1, G2, \cdots, G_9]$.

(g) Per-setp alignment.

Figure 11: RFWash evaluation for different gesture sequence lengths.

RFWash is trained/tested on sequences with lengths of one, five and ten seconds, respectively.

Results show that RFWash trained on 10 second sequences achieves GER of 7.41% which translates to a mere 0.45 substitutions, deletions or insertions to make the sequence match the ground truth per maximum number of 6 observed gestures in the sequence. Table 4 shows that the average(μ) and maximum (max) number of gestures in each sequence length along with mean and median GER. The median is zero as more than 75% of the sequences are correct with GER of zero. The table also shows that the GER decreases for larger sequences. This is because the number of errors (i.e., the numerator in GER equation) is relatively constant even as the sequence length increases. We observe that edits are usually required at the end of the sequence. We hypothesize that a gesture at the end of the sequence contains relatively fewer RD frames than other gestures due to the overlap with the next sequence. On the other hand, EMR is fairly constant across the different sequence lengths, as shown in Fig. 11c.

Table 4: The gesture recognition accuracy of RFWash

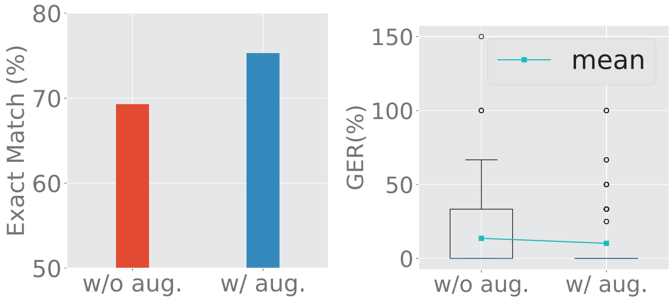
sequence length	gestures/sequence	mean GER	median GER
1s	μ:1, max:2	16.01%	0%
5s	μ:2, max:4	11.01%	0%
10s	μ:4, max:6	7.41%	0%

Recall that HCW are required to perform hand rub procedure (i.e. the nine steps) for at least 20s. We evaluate RFwash in capturing the timing of the whole handrub procedure. As the user approaches the radar, performs the procedure then walk away or stays stationary, we report the procedure time as the time of between two consecutive G_{No}. The results in Fig. 11d shows that procedure time can be estimated with very high accuracy with median error of 0s regardless of the sequence length used for training. Also, we calculated the per-step (i.e. per gesture) timing error in Fig. 11e. In general, it shows that the per-step timing error is larger than the procedure timing error. The medians absolute errors in per-step case are 0.49s, 1.17s and 1.88s for different sequence lengths of 2s, 5s and 10s; respectively.

The reason behind better procedure alignment compared to per-step alignment is that G_{No} pattern is highly distinguishable from the rest of the gestures. This makes accurate identification of the whole procedure boundaries possible even when using large sequence length (10s). On the other hand, back-to-back steps within the procedure may show similar patterns specially for mirrored gestures. Fig. 11f and 11g qualitatively compare the alignment performance for procedure and step levels; respectively. Fig. 11g illustrates per-step alignment accuracy degrade as we increase the training sequence length. Note that as the gestures are following each other in back-to-back manner, timing error in one gesture will contribute equally to the neighbouring gestures.

5.1.1 *The Impact of Data Augmentation.*

We investigate the impact of data augmentation introduced in Sec. 4.2.2 on the performance of RFWash. For the benefit of space, we show the results of sequences of 5s as other sequence lengths show similar patterns. Fig. 12 shows



(a) Exact match rates.

(b) Gesture error rates.

Figure 12: The impact of data augmentation.

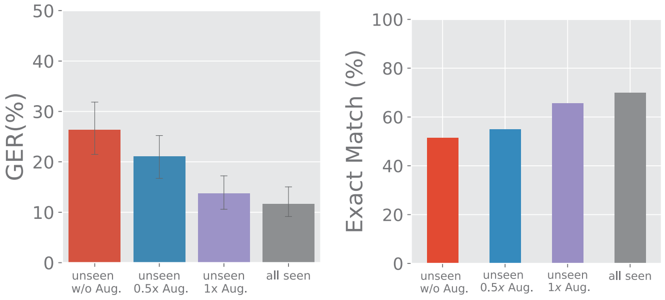
that data augmentation improves the performance of RFWash significantly. Specifically, it increases the EMR from 69.2 % to 75% and reduces mean GER from 13.6% to 11.01%. More importantly, the box plot in Fig. 12b shows that the majority of gesture sequence matches the ground truth, i.e., GER = 0.

5.2 Unseen Domains

We evaluate RFWash performance in unseen domains. This includes unseen sequence lengths, unseen gesture sequences and unseen subjects. Recall that RFWash predicts unseen gestures as G_{NO} and a misclassification of G_{NO} within a sequence contributes to the Gesture Error Rate (GER) for that particular sequence. Thus, quantitative results in Sec. 5.1 cover "unseen gestures".

5.2.1 Unseen Sequence Length. RFWash accepts radar signals of variable lengths as valid inputs (see Sec. 4.2.2). We evaluate the performance of RFWash with input sequence lengths that are different from those in the training data. The ability to classify variable length sequence is an advantage. For example, short segments have smaller latency and are preferable in scenarios where quick user feedback is needed. Longer sequences have better recognition performance and may be preferable for HH compliance audits which can be performed offline.

Fig. 15a, 15b and 15c show that unseen sequence length has negative impact on the performance of RFWash. However, the negative impact can be reduced by data augmentation, achieving significant improvements for all metrics. For example, for 15s sequence length, data augmentation improves RFWash performance by more than 2.8x, 4x and 12x, for timing estimation error, GER and EMR, respectively. We investigate this in more detail for one specific segment shown in Fig. 16. BiLSTM posteriors for three different sequence lengths of 3.1s, 6.25s and 12.5s are shown. Note that only the 6.25s sequence was included in training data. The predicted gesture sequence (G_2, G_3, G_4, G_5) for the previously seen 6s sequence is correct, both with and without data augmentation (see Fig. 16b). However, data augmentation produces significantly better temporal alignment. For unseen sequence lengths (Fig. 16a and 16c), data augmentation produces significantly better GER and temporal alignment. For example, G_2 is a false positive in Fig. 16a, while G_2, G_3 and G_7 are false negatives in Fig. 16c for RFWash without data augmentation. RFWash with data augmentation hasn't produced any incorrect predictions here. We have observed similar behavior for many other segments of signals.



(a) Gesture error rates.

(b) Exact match rates.

Figure 13: The impact of Unseen Gesture Sequences. "unseen" bars report performance when testing on gestures sequences performed in order not seen during training.

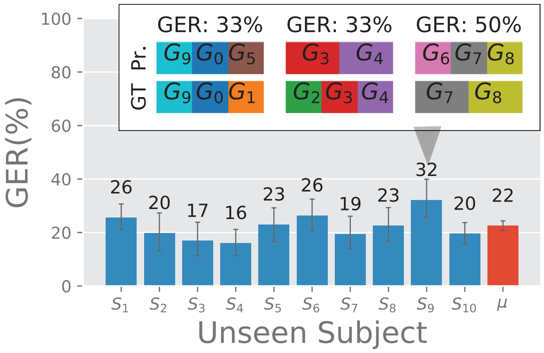
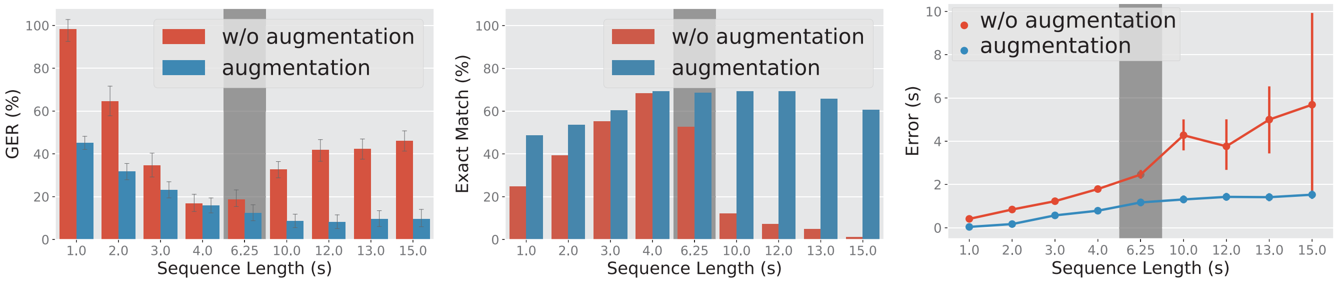


Figure 14: Impact of unseen subjects

5.2.2 Unseen Gesture Sequences. We evaluate the performance of RFWash on recognizing unseen gesture sequences. It is important RFWash performs well on unseen sequences as a HCW may follow any order of gestures (see Fig. 4 for an example), and it is difficult to collect data for all possible gesture sequences. Recall that RFWash was trained on sequences $G_1 \rightarrow G_9$ (see Sec. 5). To test RFWash performance on unseen gesture sequences, we collected additional data with hand washing gestures performed in random orders (for example, $G_9 \rightarrow G_8 \cdots \rightarrow G_1$). The additional samples collected from 4 subjects with 1-3 sessions/subject making a total of 8 sessions. Fig. 13 shows that RFWash with data augmentation performs well on unseen gesture sequences. Specifically, using full augmented dataset (1x Aug.) reduces GER by 14% compared to w/o Aug. case. This performance slightly worse (4% difference in GER) compared to testing on previously seen sequences ("all seen"). Training with half the augmented dataset (0.5x Aug.) decreases mean GER by less than 5%. This suggests that augmentation allows RFWash to close the gap on unseen sequences and, reach the performance that is close to the previously seen sequences.

5.2.3 Unseen Subjects. RFWash focuses on HH tracking for HCWs where acquiring samples from new subjects can be beneficial for identification (Sec. 7). In order to understand how the system performs for unseen subjects, we conducted a standard leave-one-subject-out cross validation [39] using 5s training sequences. Fig.14 shows that GER for unseen subjects is on average 22%, which is 11%

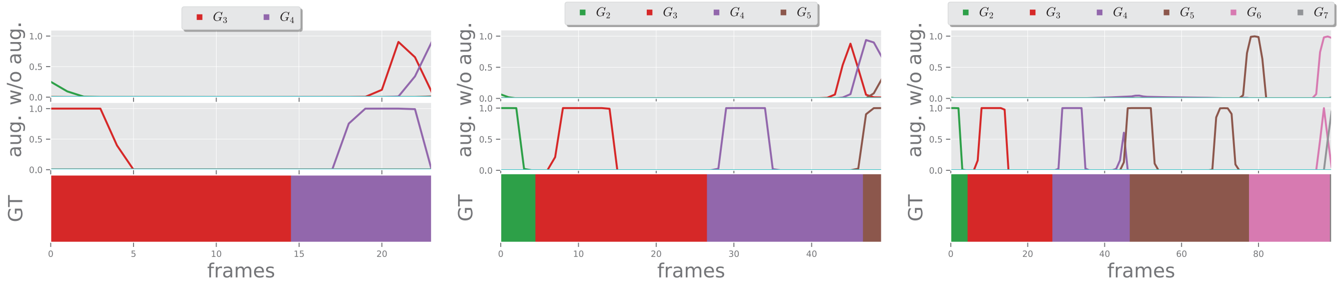


(a) Gesture error rates.

(b) Exact match rate.

(c) Timing

Figure 15: Impact of unseen sequence length. Vertical shaded areas highlights the sequence length used in the training.



(a) Alignment for unseen sequence length of 3.1s.

(b) Alignment for reference (seen) sequence length of 6.2s. We use this segment length for training.

(c) Alignment for unseen sequence length of 12.45s

Figure 16: Temporal HH gesture alignment. GT: Ground truth. aug: with data augmentation. w/o: without data augmentation

higher compared to training on all data in (Fig. 11b). For a 3-gesture sequence, this is equal to an average of 0.66 gesture edit to get an exact match on the ground truth. RFWash achieved the average EMR of 63.9%, with 87% of the sequences either matching or being a single gesture away from the ground truth. We have investigated subject S_9 in more detail as it shows a higher error with a mean GER of 32%. We found that errors corresponded mostly to sequences containing gestures G_2 and/or G_3 suggesting that confusion may be caused by subject's personal differences in performing those gestures. Nevertheless, without explicitly suppressing subjects impact using adversarial methods (as in [41]), 80% of the subject's sequences deviate from the ground truth by at most a single gesture. We note that while 32% GER might seem high, the inset plot in Fig 14 shows a few examples of sequences performed by S_9 where the prediction is off by one gesture and GER error is 33-50%.

5.3 LSTM vs BiLSTM

Table 5: LSTM vs BiLSTM

Model	2s		5s		10s	
	GER	EMR	GER	EMR	GER	EMR
(LSTM)	20.7%	69.3%	18%	54%	17.7%	44.36%
(BiLSTM)	**16%**	**76%**	**11%**	**75.28%**	**7.41%**	**76.69%**

We investigate the impact of different RFWash deep learning model parameters. Specifically, we compare the performance of CNN + LSTM and CNN + BiLSTM in Table 5, which shows that the proposed deep learning model with BiLSTM consistently performs better than that with LSTM, and the performance gap increases with the sequence length.

5.4 Comparison with Fully Supervised DL Models

In this section, we compare the performance of RFWash model to state-of-the-art supervised deep learning based approaches including C3D [36] and DeepSoli [39]. C3D has a 3D CNN that outperforms 2D CNN in large scale vision-based gesture recognition [17]. DeepSoli employs deep Convolutional Recurrent model to recognize finger gestures and operates on mmWave Range Doppler measurements of Google's Soli sensor [39] C3D and DeepSoli are trained on **manually segmented** RD frames and tested on continuous HH gesture stream using auto segmentation of a sliding window length of 8 RD frames and an overlap of 7 samples. To improve the accuracy, prediction pooling (p) is applied to C3D and DeepSoli by summing up softmax activation and using the average activation for gesture prediction [39]. RFWash is trained on continuous gesture stream data as usual with sequence length set to 2s.

Table 6 shows that RFWash outperforms the alternative approaches with an overall accuracy of 85%, which is 7% and 20% higher than C3D(p) and DeepSoli (p); respectively. We note the poor performance of DeepSoli is probably due to the low sampling rates of TI mmWave radar sensors, which is significantly lower than those of Soli radar sensors. RFWash achieves highest accuracy for most gestures except for G_2, G_3 and G_8. Furthermore, it has the additional advantage of "weak" supervision, i.e., without the requirement of intensive manual per RD frame labelling.RFWash accuracy is 4% to 9% lower than RGB (89%[20]) and depth (94%[44]) camera systems. Such systems leverage the much shorter wavelength [9] to capture exact hands shape for enhanced recognition. On the other hand, the "substantial ethical and privacy" concerns prevent using them in clinical areas as experienced by [5] .

Table 6: Recognition accuracy for different deep models

Gesture	C3D	C3D(p)	DSoli	DSoli(p)	RFWash
G_{No}	92.5%	95.1%	92.5%	94.6%	**97.8%**
G_1	69.3%	72.4%	41.6%	40.9%	**92.1%**
G_2	68.3%	76.7%	88%	**89.8%**	85.4%
G_3	84.2%	**92.9%**	76.1%	77.8	87.2%
G_4	82.2%	83.9%	77.5%	80.8%	**89.7%**
G_5	84.4%	86.4%	33.1%	34.9%	**87.7%**
G_6	48.2%	43.7%	20.3%	18.8%	**71.4%**
G_7	70.4%	69.3%	74.1%	77.2%	**79.6%**
G_8	56.7%	66.3%	76.6%	**80.6%**	77.7%
G_9	66.1%	75.9%	42%	42.1%	**84.2%**
Accuracy	73.33%	77%	63.15%	64.79%	**85%**

6 RELATED WORK

HH Monitoring Automated solutions for HH monitoring range from simple product consumption [6] and dispenser activations monitoring [15, 24] to RFID badge systems [29] for identifying the HCWs who perform HH. Also, Cameras [20] were used to enable better tracking dispenser usage. A common limitation of these approaches is that the actual HH technique is not monitored. Previous works for monitoring HH technique either employed RGB [20] or depth [44] cameras which raises privacy concerns [5, 12] in healthcare environments, or wearable sensors [16] that may cause transmission of health care-associated pathogens. RFWash, to the best of our knowledge, is the first contact-free HH technique monitoring with a potential to work inside healthcare facilities without compromising privacy.

RF Gesture Recognition Gesture recognition is a very active research field and many works applied RF-sensing to gesture recognition applications. These include WiFi-based [4, 25, 32, 37], RFID [41, 42, 45] and mmWave radar [10, 33, 39] systems. Despite the extensive treatment of the topic, little attention was paid to recognizing naturally-performed gestures that do not include pauses. This limitation continues to have its negative impact on the existing RF sensing applications such as sign language recognition [22]. Another example is WiMu [37] that successfully manages to recognize simultaneous multi-user gestures but requires user to take brief pauses before and after the gestures to enable segmentation [37]. We presented our attempt to address this problem by introducing a model that can learn on unsegmented contiguously performed gestures. We hope the results of this work spark interest of research community, and in the near future, other follow-up works can utilize unsegmented RF data streams for other sensing applications.

CTC-based Gesture Recognition A number of research works [27, 43] employed CTC-based architecture for gesture recognition applications. Most notable of which is Nvidia vision-based system [27] that fuses depth, RGB and IR camera measurements to recognize a driver's hand gestures. In these approaches, gestures are segmentable (i.e., pauses exist between gestures) and training samples are pre-segmented to contain data for one gesture only and frames for "no gesture" [27]. The role of CTC is to fine tune the predictions by locating the gesture nucleus. In our case, due to the difficulty of pre-segmenting back-to-back gestures especially from RF measurements, we proposed *training on unsegmented gesture sequence* (e.g, a number of gestures within a training segment). Such critical difference raises a challenge when training on long segments and was addressed by a novel order-preserving augmentation technique that regularizes the training process and, hence, enables learning with "weak" (less) supervision.

7 DISCUSSION AND FUTURE WORK

There is a big room for improving the current implementation of RFWash. In our design, we focused mainly on the key technical challenge of recognizing back-to-back gestures. The following are the key areas for improvement:

HH tracking inside wards: requiring the user to be stationary while facing the radar is a limitation of the current system. The prototype in its current state, however, can be employed in automated in-ward technique compliance check scenarios, i.e., a once per daily HH rub technique compliance check. Currently we are extending RFWash in two directions. First, tracking gestures of walking users/HCWs. This would require separating micro Doppler motions (i.e. hands) from the bulk Doppler (i.e. torso) using Doppler decomposition techniques [26, 30]. Second, collecting data inside wards. Limited by the ethical clearance of this research, RF data was collected only from general public in our campus. *Clinicians* may perform hand rub techniques at faster pace and interference in healthcare environment can be more challenging. Using a network of sensors in this case be investigated [12].

Accountability (HCW identification): Gesture motions pattern has been shown to be unique identifier of user [32]. This can complement RFwash and associate HH compliance with individual subjects. Identification can be done using a subsequent model trained to infer subject from predicted gesture. Following the architecture of [32], our preliminary results show that we can identify the subject who performed G_2 with an average accuracy of 94.12%.

HH tracking for general public: Maintaining good HH is one of the most effective ways to slow the spread of COVID-19. The cost of IWR1443 mmWave sensor in our prototype is less than USD 14 when bought in bulk (e.g., 1,000 units) and the form factor is small (see Figure 1). We are considering to deploy RFWash in local schools to help children maintain good HH during the challenging time of COVID-19 pandemic and beyond.

8 CONCLUSION

We introduced RFWash; an RF-based system for contact-free monitoring of healthcare workers performing Hand Hygiene techniques. The novelty of the work is two-fold. First, we fill the gap in HH technique monitoring research by introducing the first device-free system that is privacy preserving. Second, we introduce a deep model capable of recognizing back-to-back gestures that are not trivially separable. RFWash was implemented using embedded mmWave sensor and evaluated in a real-world environment. The promising results encourage us to further expand this work to collecting data at a larger scale in clinical facilities.

REFERENCES

[1] 2009. WHO Guidelines on Hand Hygiene in Health Care. Published by World Health Organisation. Retrieved from: whqlibdoc.who.int/publications/009.pdf. (2009).

[2] 2016. Healthcare-associated infections: data and statistics. Published by Center for Disease Control and Prevention. (2016).

[3] 2020 (accessed July 1, 2020). Ti IWR1443BOOST. "https://www.ti.com/tool/IWR1443BOOST". (2020 (accessed July 1, 2020)).

[4] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. 2015. Wigest: A ubiquitous wifi-based gesture recognition system. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1472–1480.

[5] Sari Awwad, Sanjay Tarvade, Massimo Piccardi, and David J Gattas. 2019. The use of privacy-protected computer vision to measure the quality of healthcare worker hand hygiene. *International Journal for Quality in Health Care* 31, 1 (2019), 36–42.

[6] John M Boyce. 2011. Measuring healthcare worker hand hygiene activity: current practices and emerging technologies. *Infection control and hospital epidemiology* 32, 10 (2011), 1016–1028.

[7] Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, and Li Fei-Fei. 2018. Privacy-Preserving Action Recognition for Smart Hospitals using Low-Resolution Depth Images. *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018* (2018).

[8] I. R. Daniels and B. I. Rees. 1999. Handwashing: simple, but effective. *Annals of the Royal College of Surgeons of England* 81, 2 (Mar 1999), 117–118. https://www.ncbi.nlm.nih.gov/pubmed/10364970 10364970[pmid].

[9] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. 2020. In-Home Daily-Life Captioning Using Radio Signals. In *European Conference on Computer Vision (ECCV) 2020.*

[10] Piyali Goswami, Sandeep Rao, Sachin Bharadwaj, and Amanda Nguyen. 2019. Real-time multi-gesture recognition using 77 GHz FMCW MIMO single chip radar. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–4.

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 369–376.

[12] Albert Haque, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, et al. 2017. Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance. *Proceedings of Machine Learning Research* (18–19 Aug 2017).

[13] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. 2020. mmVib: micrometer-level vibration measurement with mmwave radar. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[14] Yen Lee Angela Kwok, Michelle Callard, and Mary-Louise McLaws. 2015. An automated hand hygiene training system improves hand hygiene technique but not compliance. *American journal of infection control* 43, 8 (2015), 821–825.

[15] Yen Lee Angela Kwok, Craig P Juergens, and Mary-Louise McLaws. 2016. Automated hand hygiene auditing with and without an intervention. *American journal of infection control* 44, 12 (2016), 1475–1480.

[16] Hong Li, Shishir Chawla, Richard Li, Sumeet Jain, Gregory D Abowd, Thad Starner, Cheng Zhang, and Thomas Plotz. 2018. Wristwash: towards automatic handwashing assessment using a wrist-worn device. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. 132–139.

[17] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. 2016. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 25–30.

[18] Bingbin Liu, Michelle Guo, Edward Chou, Rishab Mehra, Serena Yeung, N Lance Downing, Francesca Salipur, Jeffrey Jopling, Brandi Campbell, Kayla Deru, et al. 2018. 3D Point Cloud-Based Visual Prediction of ICU Mobility Care Activities. In *Machine Learning for Healthcare Conference*. 17–29.

[19] Hu Liu, Sheng Jin, and Changshui Zhang. 2018. Connectionist temporal classification with maximum entropy regularization. In *Advances in Neural Information Processing Systems*. 831–841.

[20] David Fernández Llorca, Ignacio Parra, Miguel Ángel Sotelo, and Gerard Lacey. 2011. A vision-based system for automatic hand washing quality assessment. *Machine Vision and Applications* 22, 2 (2011), 219–234.

[21] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A. Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through Smoke: Robust Indoor Mapping with Low-Cost MmWave Radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services (MobiSys '20)*. 14–27.

[22] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign language recognition using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 23.

[23] AR Marra and MB Edmond. 2014. New technologies to monitor healthcare worker hand hygiene. *Clinical Microbiology and Infection* 20, 1 (2014), 29–33.

[24] Maryanne McGuckin and John Govednik. 2015. A review of electronic hand hygiene monitoring: considerations for hospital management in data collection, healthcare worker supervision, and patient perception. *Journal of Healthcare Management* 60, 5 (2015), 348–361.

[25] Pedro Melgarejo, Xinyu Zhang, Parameswaran Ramanathan, and David Chu. 2014. Leveraging directional antenna capabilities for fine-grained gesture recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 541–551.

[26] Marco Mercuri, Ilde Rosa Lorato, Yao-Hong Liu, Fokko Wieringa, Chris Van Hoof, and Tom Torfs. 2019. Vital-sign monitoring and spatial tracking of multiple people using a contactless radar-based sensor. *Nature Electronics* 2, 6 (2019), 252–262.

[27] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4207–4215.

[28] Avishek Patra, Philipp Geuer, Andrea Munari, and Petri Mähönen. 2018. mm-Wave Radar Based Gesture Recognition: Development and Evaluation of a Low-Power, Low-Complexity System. In *Proceedings of the 2nd ACM Workshop on Millimeter Wave Networks and Sensing Systems*. 51–56.

[29] Lisa L Pineles, Daniel J Morgan, Heather M Limper, Stephen G Weber, Kerri A Thom, Eli N Perencevich, Anthony D Harris, and Emily Landon. 2014. Accuracy of a radiofrequency identification (RFID) badge system to monitor hand hygiene behavior during routine clinical activities. *American journal of infection control* 42, 2 (2014), 144–147.

[30] Xingshuai Qiao, Tao Shan, Ran Tao, Xia Bai, and Juan Zhao. 2019. Separation of human micro-Doppler signals based on short-time fractional Fourier transform. *IEEE Sensors Journal* 19, 24 (2019), 12205–12216.

[31] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[32] Muhammad Shahzad and Shaohu Zhang. 2018. Augmenting User Identification with WiFi Based Gesture Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 134.

[33] Karly A Smith, Clément Csech, David Murdoch, and George Shaker. 2018. Gesture recognition using mm-wave sensor for human-car interface. *IEEE Sensors Letters* 2, 2 (2018), 1–4.

[34] Jocelyn A Srigley, Colin D Furness, G Ross Baker, and Michael Gardam. 2014. Quantification of the Hawthorne effect in hand hygiene compliance monitoring using an electronic monitoring system: a retrospective cohort study. *BMJ Qual Saf* 23, 12 (2014), 974–980.

[35] Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. 2018. RF-based fall monitoring using convolutional neural networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[37] 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*. ACM, 401–413.

[38] Aditya Virmani and Muhammad Shahzad. 2017. Position and orientation agnostic gesture recognition using wifi. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 252–264.

[39] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 851–860.

[40] Xin Yang, Jian Liu, Yingying Chen, Xiaonan Guo, and Yucheng Xie. 2020. MU-ID: Multi-user Identification Through Gaits Using Millimeter Wave Radios. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2589–2598.

[41] 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*. 298–310.

[42] Shigeng Zhang, Chengwei Yang, Xiaoyan Kui, Jianxin Wang, Xuan Liu, and Song Guo. 2019. ReActor: Real-time and Accurate Contactless Gesture Recognition with RFID. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[43] Zhenyuan Zhang, Zengshan Tian, and Mu Zhou. 2018. Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor. *IEEE Sensors Journal* 18, 8 (2018), 3278–3289.

[44] Henry Zhong, Salil S Kanhere, and Chun Tung Chou. 2016. WashInDepth: Lightweight Hand Wash Monitor Using Depth Sensor. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, 28–37.

[45] Y. Zou, J. Xiao, J. Han, K. Wu, Y. Li, and L. M. Ni. 2017. GRfid: A Device-Free RFID-Based Gesture Recognition System. *IEEE Transactions on Mobile Computing* 16, 2 (Feb 2017), 381–393. https://doi.org/10.1109/TMC.2016.2549518