IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 17, 2019, accepted December 22, 2019, date of publication January 1, 2020, date of current version January 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963440

Ambient Light Based Hand Gesture Recognition Enabled by Recurrent Neural Network

HAIHAN DUAN[1], MIAO HUANG[1], YANBING YANG[1,2], JIE HAO[3], AND LIANGYIN CHEN[1,2]

[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]The Institute for Industrial Internet Research, Sichuan University, Chengdu 610065, China
[3]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Corresponding authors: Yanbing Yang (yangyanbing@scu.edu.cn) and Liangyin Chen (chenliangyin_scu@qq.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant YJ201868, in part by the National Natural Science Foundation of China under Grant 61902267 and Grant 61602242, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20160807.

ABSTRACT As an essential requirement of pervasive smart devices, free hand gestural input considered as necessary for user interactions has attracted lots of research attention for nearly decades. Nevertheless, existing proposals heavily rely on either expensive pre-deployed equipment or user on-body sensors, thus confine their application scenarios. In this paper, we propose a novel hand gesture recognition system which purely relies on ubiquitous ambient light and low-cost photodiodes. The proposed system does not need any modification to existing lighting infrastructure. While without complex signal pre-processing for modulated light, very low-cost photodiodes and processors can capture and process the light variations caused by hand gesture. To produce accurate hand gesture recognition, we design efficient algorithms based on recurrent neural network to process sensing data collected by a photodiode array. We implement a prototype consisting of an array of 8 photodiodes and extensive experiments demonstrate that the proposed solution can achieve a very high overall recognition accuracy of 99.31%.

INDEX TERMS Visible light sensing, recurrent neural network, hand gesture recognition.

I. INTRODUCTION

To seek applicable hand gesture recognition solutions, a lot of researchers from both academic and industrial communities devote to this field, and many methods are proposed. In early years, a camera is employed to capture and recognize hand gesture since it imitates our eyes to ''see'' the gesture [1]–[3]. However, using the camera may cause security and privacy issue, so other invisible signals, e.g., ultrasonic [4], [5] and radio-frequency [6], [7], are used to sense our hand gesture. Obviously, ultrasonic based solutions confine their application scenarios since they need to deploy both supersonic source and receiver which are not so common in existing infrastructure. To release the load of deploying signal source, existing RF signals, e.g., indoor WiFi signals, are hence well explored in many proposals [6]–[10]. **Nevertheless, existing radio-frequency enabled solutions: on the one hand require delicate radio-frequency source deployment to guarantee recognition accuracy, on the other hand are prone to environment variation and electromagnetic interference. Therefore, pervasive visible light is considered as an applicable and promising sensing medium for indoor gesture recognition.**

Several seminal solutions have been proposed to recognize hand gestures through visible light [11]–[13]. Especially, Aili in [11] is able to accomplish arbitrary 3D hand gesture reconstruction with delicately modulated light which is generated by an LED plane with 288 LED chips and captured by a 16 photodiodes array. To release the bottleneck of heavy deployment for both light source of LEDs and massive sensors of photodiodes, in this paper, we propose to use unmodulated ambient light as the sensing medium for hand gesture recognition. In essential, our idea is very intuitive: in any human existing indoor environment, visible light is required and our hand's shadow always casts on some surface, so sensing such shadow could allow us to recover the hand gesture. As illustrated in Fig 1, the sensing value generated by the photodiode array indeed changed under different

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Tucci.

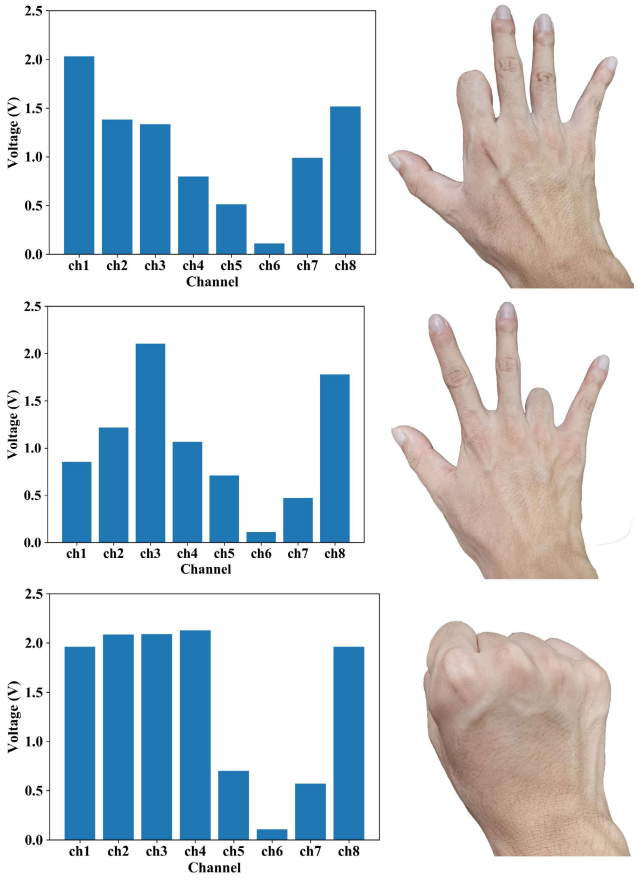


FIGURE 1. Sensing value variances under different hand gestures.

hand gestures. However, such a straightforward idea of using shadow to infer hand gesture imposes on us a big challenge: without a clear geometrical relationship between the hand and the photodiodes, only using the variance in voltage obtained by the photodiode sensor is very hard to recover the hand gesture. So we propose to utilize advanced recurrent neural network (RNN) to recognize the multi-dimensional sensed data to the desired hand gesture.

To verify our idea of using ambient light for hand gesture recognition, we build a prototype consisting of 8 low cost photodiodes to record the shadow of hand gesture. We implement the RNN-based algorithm for data processing and hence gesture recognition. Our main contributions are as follows:

- We propose a novel idea of using ambient light for hand gesture recognition, and build a prototype consisting of 8 low cost photodiodes to showcase the feasibility and effectiveness of this lightweight hand gesture recognition approach.
- **We design a data processing and RNN-based algorithm to map the multi-dimensional sensing data to form the shadow of hand gesture so as to recognize each hand gesture.**
- We conduct extensive field experiments based on the implemented prototype, and the results show its high accuracy of 99.31% for hand gesture recognition.

In the following, we first introduce the research background via briefly reviewing existing literature on hand gesture recognition in Section II. The detailed gesture recognition approach based on RNN framework is presented in Section III, followed by the extensive field experiments for evaluating its performance reported in Section IV. Finally, we conclude this paper in the last section of Section V.

II. RELATED WORK

For device-free gesture recognition, it can be divided into three categories according to the sensing medium, namely radio-frequency, acoustic and visible light signals. We here hence briefly summarize some of the important existing works to have an overview and better position our work.

A. RADIO FREQUENCY

Using radio frequency for gesture recognition is a very hot research topic recently. Many pioneers take their efforts on this field [6]–[8], [10], [14], [15]. Qifan et al. proposed and implemented Wisee [6] which can enable whole-home gesture recognition through walls depending on several WiFi access points. Following the idea of using WiFi signals, the authors in [10] built a gesture recognition system utilizing reflected signals from user's hands so as to track the motion of the hands and recognize corresponding gestures. To address the issue of identifying the user and recognizing gesture at the same time, WiID [7] explores the time series of the WiFi signal's frequency changes caused by different users when performing given gestures. Besides of WiFi signals, ambient wireless signals, such as TV transmissions [9] and mobile communication (GSM) [16], even advanced 60 GHz signals [15] were also investigated to be used for gesture recognition. Intuitively, radio frequency based gesture recognition requires pre-deployed or existing wireless sources and reception instruments, so the recognition is inevitable to be a complex system which means a relatively expensive hardware cost. On the contrast, visible light is a free and pervasive sensing medium which can be captured by very low cost photodiodes or even the illuminating LEDs [17]. Moreover, visible light enabled gesture recognition has much wider availability comparing to radio based techniques, since it has no electromagnetic interference and electromagnetic compatibility issues.

B. ACOUSTIC SIGNALS

Over the years, acoustic signals are still actively exploited in research fields of hand gesture recognition and motion sensing. Kalgaonkar and Raj [18] engineered a device that could recognize 8 one-handed gestures. SoundWave [4] recognizes coarse-level gestures including one- and two-handed gestures and walking motion, using the existing hardware embedded in a computer. By leveraging various attributes of identified 6 hand gestures, AudioGest [19] provides hundreds of control commands for possible applications. FingerIO [5] and LLAP [20] achieve millimeter-level finger tracking with mobile devices. Most of the works

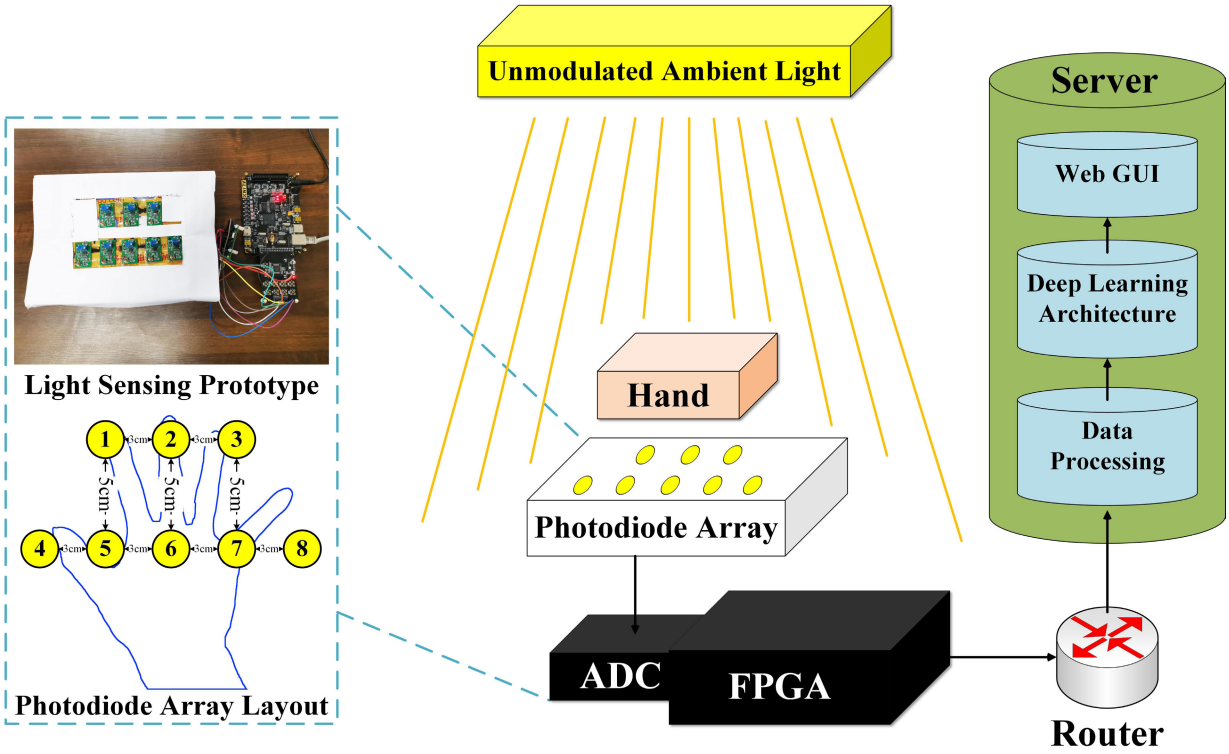


FIGURE 2. Experimental flowchart of the hand gesture recognition.

make use of the well-understood Ultrasonic Doppler techniquVs [4], [5], [18] and prefer COTS speakers and microphones embedded in commodity devices [4], [19], [20], which may raise privacy issues. Visible light-based recognition methods, however, are free from such issues.

C. VISIBLE LIGHT

For using visible light as sensing medium for gesture recognition, cameras are first considered as the popular receivers to capture and recognize gestures, and many existing works have exactly done [1]–[3]. However, using cameras for hand gesture recognition may result in privacy issues. Our work in this paper focuses on sensor based hand gesture recognition, therefore, we omit the detailed review on camera based works yet concentrate on hand gesture recognition enabled by photodiodes or light sensors. Okuli [12] is presented as an Android peripheral with two light sensors and an LED to locate a finger on a 2-D plane. Tianxing et al. propose the system LiSense [21] to perform continuous 3D human skeleton reconstruction in real time with 5 LED lights and 324 off-the-shelf photodiodes, and Aili [11] accomplishes arbitrary 3D hand gesture reconstruction with 288 LED lights and 16 photodiodes, styled as a table lamp. By modulating the LED lights with various frequencies, each photodiode is able to distinguish between different light sources so as to construct a shadow map, which in combination can be interpreted as corresponding gestures. Attaching a receiver adjacent to each of the six LED lights on the ceiling, EyeLight [22] comes up with a tracking algorithm and explores the possibility of room activity and occupancy recognition with a supervised machine learning approach. EyeLight utilizes reflected light off the floor to feed its server and a time division signaling scheme to separate different light sources. **What's the difference about our system is that we take advantages of the ubiquitous ambient light to recover a set of hand gestures which requires no modification to existing light infrastructure, and as the system does not need to modulate the light transmitter, it can be adapted in almost any environment with people.** Recent works [13], [23] make efforts to incorporate solar cells instead of photodiodes as the sensors of visible light, both of which realize hand gesture recognition through unmodulated ambient light and achieve an obvious drop in energy consumption. **Given the limited information obtained from the input of the solar panel, these systems have restricted competence in a more fine-grained and static gesture recognition, whereas our system produces good results in relatively subtle finger difference.**

III. METHODOLOGY

A. EXPERIMENTAL SETTING

The experimental flowchart of hand gesture recognition is shown in Fig. 2. A light sensing prototype consisting of 8 sensing units is implemented for capturing the intensity change of unmodulated ambient light introduced by different hand gestures. The detailed illustration of the light sensing prototype can be found in the left part of Fig. 2. **Each sensing unit consists of a low cost photodiode of Honeywell SD5421-2 and a low-end amplifier**

FIGURE 3. Seven selected hand gestures.

of STMicroelectronics LM358, so that the sensing unit can work under an illuminance level as low as 200 lux. We empirically design the layout of photodiodes, in which three sensing units are put in the first row to cover the shadow blocked by the mid three fingers and the remainder five are put in the second row. The distance between two rows is set as 5cm, and each sensors are put with a closer interval of 3cm. **The selected photodiode array layout has at least two advantages: 1) it corresponds to the shape of people's hand so that the change of each finger could be effectively captured; 2) it is easy for people to locate a roughly center position (the line between the** $2nd$ **and the** $6th$ **sensor in photodiode array layout of Fig. 2) to put their hands upon the sensors thanks to the symmetrical design. The photodiode captures the light intensity change caused by the hand placed over the prototype, and converts the incident light to tiny current which will be amplified by a transconductance amplifier built upon STMicroelectronics LM358. A Field-Programmable Gate Array (FPGA) development board with an 8-channel Analog-Digital Converter (ADC) daughter board (sampling rate** $SR_{ADC} = 4096Hz$**) is used to transform the analog voltage signal to digital data stream and passes the sensed data to a server for recognition through a router via Ethernet.** The server decodes the received data to predefined format and utilizes an effective algorithm based on deep learning [24], which would be described in Section III-C, to recognize the corresponding hand gesture. A web based graphical user interface (GUI) is implemented to interact with the server and the light sensing prototype. **Note that, by simply adding more sensing units and ADC daughter boards, our prototype can be quickly scaled to adapt two-hand gesture recognition.**

To better examine our prototype, we intuitively choose five common and practical hand gestures to mimic how we press the keyboard in air and two other gestures of open hand and fist, as shown in Fig. 3. We elaborate each gesture as follows:

One finger down: **Be supposed to the basic and common movement when we interact with a keyboard, one finger pressing down should be evaluated at first.** We graphically show them as numbers 1-5 in Fig 3.

Open hand and fist: The other two kinds of gestures that we choose are open hand and fist since they are useful and can provide more possibility of interaction.

B. DATASET

For sensing the shadow of hand gesture, eight low cost photodiodes are employed to form a sensing array to detect the ambient light changes blocked by hand as shown in Fig. 2. To be more specific, the sensing array front-end captures the light intensity variances caused by different hand gestures and converts weak light changes into voltage signals. **An 8-channel ADC daughter board which is host by an FPGA development board as detailed in Section III-A converts the analog voltage to digital data stream, and the data stream could be defined as:**

$$D = \{X^{(i)} | i = 1, 2, \ldots, (T \times SR_{ADC})\} \tag{1}$$

in which $X^{(i)} \subset \mathbb{X}^n$ represents the *ith* n-dimensional vector where n denotes the number of photodiodes, T represents the duration of data collection, and SR_{ADC} represents the sampling rate of the ADC. If we set the sampling rate of each data frame as SR_{frame}, the dataset could be represented as:

$$\mathbb{D} = \{f^{(i)}, l^{(i)} | i = 1, 2, \ldots, \frac{T \times SR_{ADC}}{SR_{frame}}\} \tag{2}$$

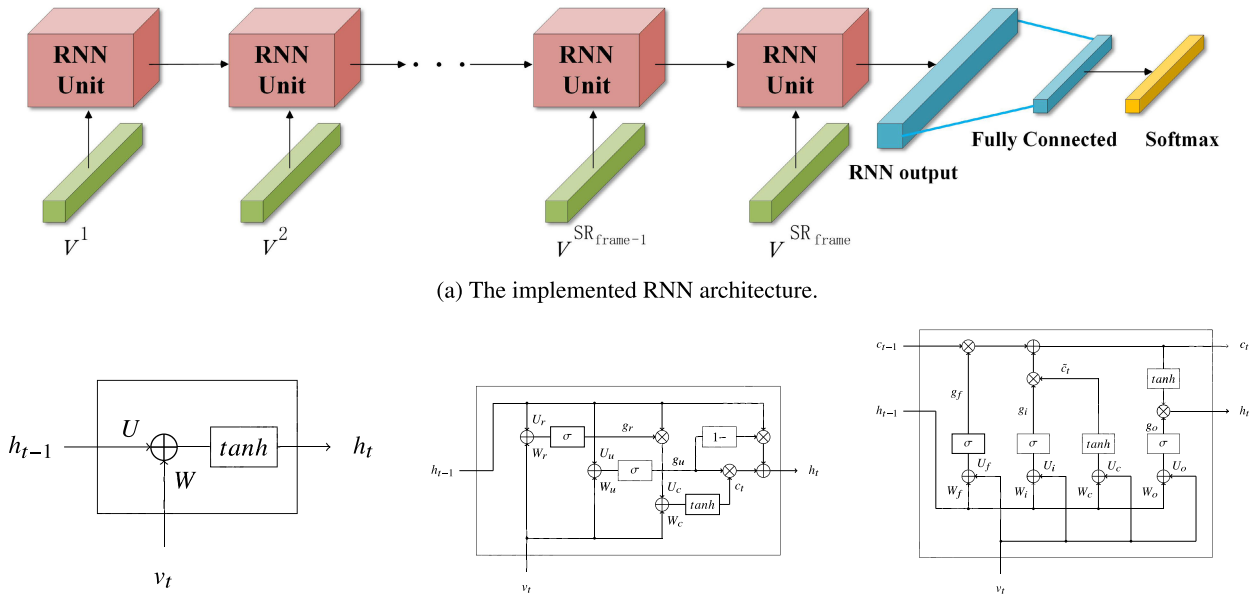
in which $f^{(i)} \subset \mathbb{F}^{n \times SR_{frame}}$ represents *ith* data frame which has dimension of $n \times SR_{frame}$, and $l^{(i)} \subset \mathbb{G}^c$ denotes the label of $f^{(i)}$ where $\mathbb{G}^c$ is the one-hot encoding of c classes of gestures we selected to recognize. Therefore, each data frame is a temporal sequence consisted of n-dimensional vectors, which could be defined as:

$$f^{(i)} = \{v^t | t = 1, 2, \ldots, SR_{frame}\} \tag{3}$$

in which $v^t \subset \mathbb{V}^n$ represents n-dimensional vector collected at time t of the sampling window. **During data collection, the users need to put his/her hand upon the photodiode array to make the shadow of five fingers cover all sensors, and align his/her middle finger with the symmetrical line of photodiode array (the line between the** $2nd$ **and the** $6th$ **sensor in photodiode array layout of Fig. 2) to roughly fix the position of hand.** Then the user keep their hand static for several seconds when the server records the sensing data and labels the data as a specific gesture. More details of data collection in different experimental settings will be discussed in Section IV.

C. RECURRENT NEURAL NETWORK ARCHITECTURE

To recover and recognize the hand gesture from the 8-dimensional vector generated by our prototype as aforementioned in Section III-A, we design an effective recognition algorithm based on the advanced deep learning. Deep learning is a category of machine learning technology which utilizes artificial neural network as computational model with backpropagation algorithm to update parameters [24]. Therein, the recurrent neural network (RNN), a category of deep learning method, is very powerful in processing sequence of time steps since it can retain information about the history of all past elements of the sequence [24]. As mentioned in Section III-B, the data frame of our task is a temporal sequence which is highly suitable to be processed by RNN. Therefore, we implement an RNN framework to recognize hand gestures as shown in Fig. 4a, with a fully connected layer following with a Softmax activation layer as the output layer.



(a) The implemented RNN architecture.

(b) SimpleRNN unit.

(c) GRU unit.

(d) LSTM unit.

FIGURE 4. Illustration of the RNN architecture and different RNN units.

To satisfy the real-time requirement, we use only one layer of RNN with 128-dimensional vector for each RNN unit. Furthermore, in order to effectively recognize hand gestures, we need to select a category of RNN unit which can achieve a high accuracy with low latency. However, with the development of natural language processing (NLP) tasks in recent years, the RNN unit shows large diversity [25]. Since we cannot try all designs of RNN unit, we empirically select three extremely classical and representative types as candidates as shown in the bottom part of Fig. 4, namely Simple Recurrent Neural Network (SimpleRNN) [26], Gated Recurrent Unit (GRU) [27], and Long Short-Term Memory (LSTM) [28], which are also the fundamental forms of most novel RNN units. Different RNN units have different attributes which may result in correspondingly different performance. Here we first introduce their architectures in theory to better fit our application of hand gesture recognition.

The SimpleRNN unit is the original form of RNN which only contains two parameter matrices W and U as shown in Fig. 4b. The formula of a SimpleRNN unit is defined as:

$$h_t = tanh(Wv^{(t)} + Uh_{t-1} + b) \tag{4}$$

in which h_t represents the output vector of time step t, the W and H denote two parameter matrices, b represents the bias vector, and *tanh* means the tanh activation function. Because of the simplicity of SimpleRNN unit, the vanishing gradient problem becomes more serious [29]. Even worse, when the RNN model represents long-term dependencies, the gradients from back-propagation could be exponentially small. Moreover, the information of long-term dependency is easily disturbed by short-term relationship caused by small signal fluctuation. **Although the unstable ambient light variance and tiny hand tremor may slightly degrade the performance of the SimpleRNN unit in our ambient light based hand gesture recognition application, it is still worth to explore the ability of SimpleRNN given its simplicity of implementation and low computational complexity.**

Favored as an improvement of SimpleRNN unit, two gates for determining how much new information and historical features should be retained in output are added in GRU as shown in Fig. 4c, which is also known as a category of gated RNN [27]. The formula of GRU unit can be represented as:

$$g_u = \sigma(W_u v_t + U_u h_{t-1} + b_u) \tag{5}$$
$$g_r = \sigma(W_r v_t + U_r h_{t-1} + b_r) \tag{6}$$
$$c_t = tanh(W_c v_t + U_c(g_r \otimes h_{t-1}) + b_c) \tag{7}$$
$$h_t = g_u \otimes c_t + (1 - g_u) \otimes h_{t-1} \tag{8}$$

in which g_u and g_r denote update gate and reset gate respectively, σ denotes sigmoid function which returns a value from 0 to 1, and $\otimes$ operation is element-wise multiply. Similar to the leaky unit [30], the intuition of the gated RNN is to control gradients without diminishing or explosion. The leaky unit needs to manually select coefficients for the purpose of parametric connection weights, while the gated RNN could vary the connection weights dynamically in every time step. For GRU unit, the update gate g_u and reset gate g_r can independently filter the information of state vector. The update gate g_u can selectively choose the new state vector c_t to update the target vector h_t. And the reset gate g_r controls the historical state vector h_{t-1} through evaluating the relationship between historical state vector h_{t-1} and cell state vector h_t. **Due to the two additional gates, the GRU unit has threefold parameters compared with SimpleRNN unit, so it can**

get a better recognition performance with slightly higher computational cost and complexity.

The LSTM unit is another form of gated RNN as shown in Fig. 4d, which was proposed with the intuition of changing connection weights dynamically according to the context before the GRU. The expression of LSTM is defined as:

$$g_o = \sigma(W_o v_t + U_o h_{t-1} + b_o) \quad (9)$$
$$g_f = \sigma(W_f v_t + U_f h_{t-1} + b_f) \quad (10)$$
$$g_i = \sigma(W_i v_t + U_i h_{t-1} + b_r) \quad (11)$$
$$\tilde{c}_t = tanh(W_c v_t + U_c h_{t-1} + b_c) \quad (12)$$
$$c_t = g_i \otimes \tilde{c}_t + g_f \otimes c_{t-1} \quad (13)$$
$$h_t = g_o \otimes tanh(c_t) \quad (14)$$

in which $\tilde{c}_t$ represents a temporary state vector of output, and c_t denotes the cell state vector. The GRU unit uses reset gate g_r and update gate g_u to determine both the forget coefficient and update state vector at the same time, while the LSTM unit applies three independent gates to control the information transfer of state vectors, including an input gate g_i, an output gate g_o and a forget gate g_f. The forget gate g_f controls how much the historical cell state vector c_{t-1} should affect the new cell state vector c_t. **And the input gate g_i is computed similarly to the forget gate g_f, but it also obtains the gating value of input state vector $\tilde{c}_t$.** Then, the new cell state vector c_t is calculated by element-wise adding two gating values from input gate and forget gate. Finally, the output gate g_o filters the cell state vector c_t as output vector h_t of the time step. **In fact, the LSTM is proved to be "strictly stronger" than the GRU since it can be more easy to perform unbounded counting and learn long-term dependencies compared with the simple architectures [31].** However, the LSTM unit has the largest number of parameters in selected RNN units, hence it can cause a long latency of recognition.

These three aforementioned RNN types will be evaluated one by one via our prototype to find the unit with the best balance between accuracy and latency in the following Section IV-A.

D. MODEL TRAINING

We implement the RNN framework using Keras [32] with backbone of Tensorflow [33], and train and test on a server with 4 kernel processors (Intel Core i7 7700 @3.6GHz) and an NVIDIA GeForce 1080Ti (11GB) with 16GB RAM. For the loss function, the categorical cross-entropy is preferable for multi-label classification tasks when the data is lightly unbalanced. So we choose the categorical cross-entropy as loss function, which is defined as:

$$L(l, y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}(l_{ij} * log(y_{ij})) \quad (15)$$

in which N means the total number of testing data frames, c represents the number of classes, l and y denote the true label and predicted result respectively. The goal of model training is to optimize the weight matrices (e.g., W or U) to make the loss function $L(l, y)$ reach its minimum through backpropagation algorithm. The Adam optimizer [34] is utilized with the learning rate of 1.0×10^{-4}. Learning rate decay is applied for better convergence of our model, and weight decay factor, patience and minimum learning rate are set as 0.2, 5 and 1.0×10^{-5} respectively. Moreover, we set the number of epochs as 200 with batch size of 512.

IV. EVALUATIONS

In this section, we extensively evaluate the performance of the proposed hand gesture recognition scheme under various experimental settings. Furthermore, based on the experiment results, we also discuss the potentials and limitations about this novel visible light enabled gesture recognition solution.

A. IMPACT OF DIFFERENT RNN UNITS AND SAMPLING RATE

At first, we evaluate the performance in terms of accuracy and latency under three selected RNN units,namely SimpleRNN, GRU and LSTM as analyzed in Section III-C. Four volunteers are recruited for data collection, including three males and one female. For each volunteer, 200-second sensing data per gesture are collected with right hand above the sensing units with a distance of 10cm. The dataset is randomly divided into three subsets: training set, validation set and testing set according to the percentage of 80:10:10. As for the sampling rate SR_{frame}, we configure it as 64 Hz, 128 Hz, 256 Hz and 512 Hz. Note that, to ensure the sensed data being processed in real time, the latency should be lower than $\frac{1}{SR_{frame}}$, we hence define a metric of RR representing the real-time requirement for latency evaluation. Besides, as mentioned in Section III-B, the sum of all sensing data is defined as $T \times SR_{ADC}$, in which $T = 200s$ and $SR_{ADC} = 4096Hz$, and the amount of data frame is defined as $\frac{T \times SR_{ADC}}{SR_{frame}}$, determined by SR_{frame}. Therefore, for lower sampling rate of data frame, the dataset clearly has more raw sensing data compared with higher sampling rate.

The experiment results are shown in Table. 1, where SR_{frame} denotes sampling rate of data frame, DA means raw data amount, RR represents the real-time requirement, L denotes latency, VA and TA represent accuracy of validation set and testing set respectively. As we would expect, under lower sampling rate we can get a higher recognition accuracy comparing with higher sampling rate since we can get more raw data for model training. Among three RNN units, the model with LSTM unit achieves the highest test accuracy of 0.9998 under a data frame sampling rate of 64 Hz. But as a cost of using LSTM units, the computational latency is the largest one comparing with SimpleRNN and GRU units under the same sampling rate of data frame. Specifically, the latency with LSTM units under sampling rate of 64 Hz is 0.575s, which could not satisfy the real-time requirement of 0.313s, so it can not be a candidate for our system envisioned as for realistic application. **However, the GRU with a data**

TABLE 1. Performance under different RNN units and sampling rates.

			SimpleRNN			GRU			LSTM		
$SR_{frame}(Hz)$	DA	$RR(s)$	VA	TA	$L(s)$	VA	TA	$L(s)$	VA	TA	$L(s)$
64	179200	0.0313	0.9898	0.9899	0.0313	0.9976	0.9968	0.0499	0.9998	0.9998	0.0575
128	89600	0.0625	0.9599	0.9550	0.0396	**0.9941**	**0.9931**	**0.0622**	0.9943	0.9927	0.0725
256	44800	0.1250	0.8801	0.8757	0.0528	0.9900	0.9897	0.0849	0.9650	0.9612	0.1025
512	22400	0.2500	0.8996	0.8955	0.0834	0.9228	0.9188	0.1339	0.8134	0.7973	0.1584

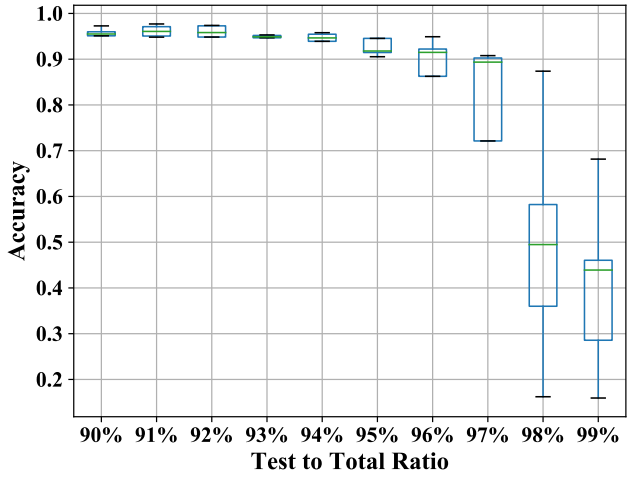


FIGURE 5. Accuracy with varying TTR from 90% to 99%.

frame sampling rate of 128 Hz achieves the best balance between accuracy and latency, i.e., a testing accuracy of 0.9931 and a latency of 0.0622s (less than the real-time requirement of 0.0625s). As a result, we employ the GRU unit as the selected RNN model with data frame sampling rate of 128 Hz for further experiments of our prototype.

B. IMPACT OF TRAINING INTENSITY

To evaluate how much training data is needed in our system for achieving a reasonable recognition accuracy, we take a substantial fraction of the labelled data as testing data and use the remaining for training the model. We so define the Test to Total Ratio (TTR) to indicate the fraction of data chosen for testing the model, and vary the TTR from 90% to 99%. As shown in Fig. 5, the average accuracy of the model is decreasing along with the increase of TTR (less training data). Besides, with higher TTR, the performance of the model illustrates an unstable status with large bias of accuracy due to deficient training limited by very little data. In contrast, the accuracy with TTR of 90% has a small deviation and closes to 0.95, which is only 4% lower than the best recognition performance of 0.9931. In fact, if we slightly lower the recognition accuracy to 95%, which is sufficient for many realistic applications, e.g., interacting a washing machine with this technology, only 6% training data could be needed. **So this lightweight ambient light based hand gesture recognition technology can be quickly deployed with extensive training.**

C. IMPACT OF HAND'S POSITION

Since the outline of the hand's shadow may change when the user changes the distance between his/her hand and the sensing units, we verify the robustness of our prototype with varying the above distance between the user's hand and the sensing units. We vary the distance ranging from 5cm to 15cm with a step of 5 cm. Fig. 6 reports the experiment results. As described in III-A, the numbers of 0 to 6 denote the seven hand gestures in this paper. As expected, the best performed height is 10cm, because our learning model is trained by the data collected under this setting. As shown its confusion matrix, under a lower position of 5cm above the sensing units, the gesture of open hand is mostly misrecognized as little finger down. It can be understood since the shadow of little finger cannot cover the 8*th* photodiode in Fig. 2 under the height of 5cm. However, the other gestures are still recognized precisely, so our current prototype is hence compatible when the user poses his/her hand slightly close to the sensing units. As for the case with far distance away the sensing units, the results are quite unsatisfied. **That is because, on the one hand the shadow becomes indistinct to be sensed by the sensing units, on the other hand the outline of hand's shadow may be larger or shift beyond the sensing range covered by our current prototype of 8 sensing units.** Fortunately, we normally interact with our appliances in at a close distance and it is quite rare to put the hand so high beyond 10cm above the devices. **In addition, it is quite easy to slightly adjust our prototype with tuning the hardware or re-training the learning model to accommodate new environmental conditions.**

D. IMPACT OF DIFFERENT USERS

Envisioned as a new interaction technology, massive users are expected, so we further evaluate the robustness of our prototype with different users because everyone has different hand shapes and finger lengths. We recruit three additional volunteers to make the seven pre-defined hand gestures above the prototype. We report the results in Fig. 7. Obviously, the recognition performance degenerates for some gestures, because it is true that our hands are highly diverse, especially for male and female. Moreover, it is hard for some users to make some gestures evaluated in this paper. For example, the first user could not raise his ring finger when his little finger is bending, so almost all cases of little finger down are recognized as the gesture of ring finger down. But for

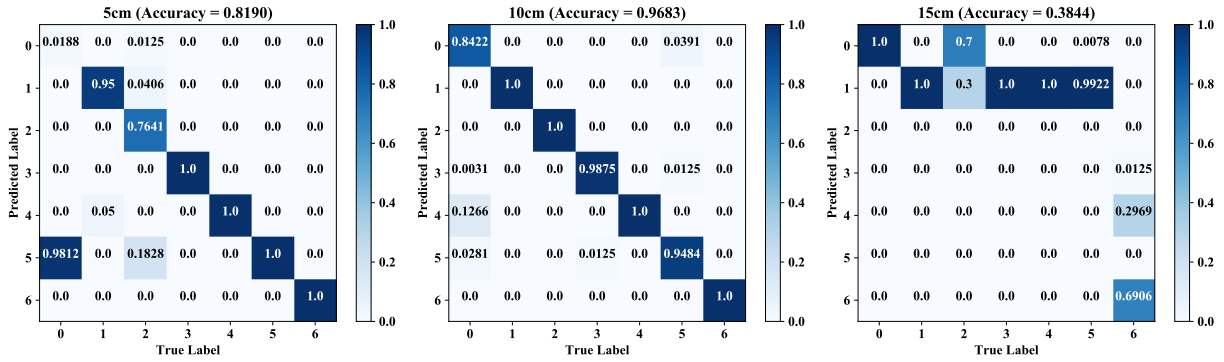


FIGURE 6. Accuracy under different heights between hand and sensing units.

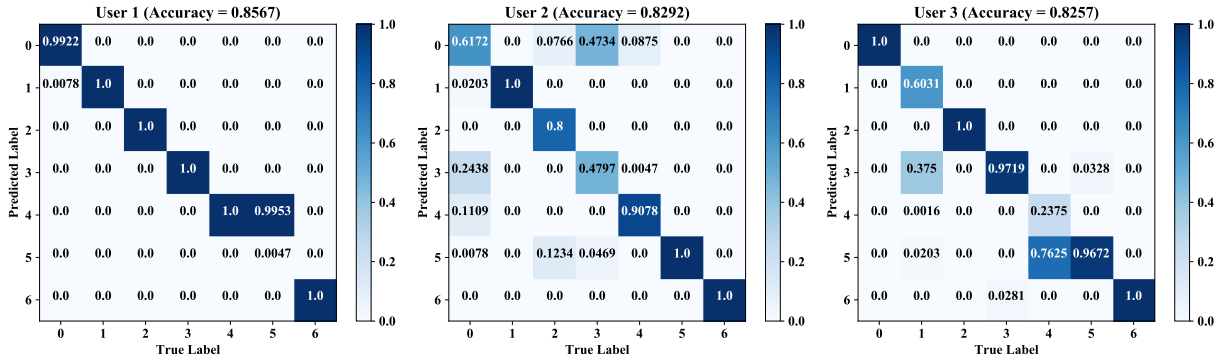


FIGURE 7. Accuracy of different users.

the other gestures, our system achieves a promising performance. For the second user, we find his middle finger could only bend to a very small degree, so the shadow variance caused by bending his middle finger is too tiny to be sensed by our prototype. Similarly, the third one needs to nip his middle finger and little finger when his ring finger is bending, which results in that the sensing units under his ring finger is covered by his little finger. Therefore, the particularity of different users is indeed making impact on the performance of our current prototype. Nevertheless, if the user could make the gestures roughly precisely, our prototype can definitely recognize their hand gestures. **In realistic application scenarios, users could often calibrate the system via simply re-training the learning model before using it to achieve the best suitability for a better service.**

E. POTENTIALS AND LIMITATIONS

Although our current prototype demonstrates promising performance in hand gesture recognition, we have also found some of its limitations and potential extensions. Therefore, we seriously discuss them in this section.

1) MULTIPLE GESTURES

As reported in aforementioned section, our prototype can currently recognize seven pre-defined gestures with a high accuracy, and it can satisfy many realistic applications, e.g., interacting with a washing machine or using it to play a virtual piano. While it is worth to largely extend the recognition ability for more gestures, for example, recognizing combinational gestures involved multiple fingers to mimic clicking two or three buttons simultaneously on a keyboard. **And we are on the way to train the learning model with more labelled gesture data and build a larger prototype with more sensing units to effectively enhance the recognition capability of our unmodulated visible light based hand gesture recognition scheme. On the other hand, our current prototype mainly focuses on the static gestures and could not support dynamic gestures. But the RNN-based method has ability to capture the changing trend of light intensity so as to infer large amplitude dynamic gestures, which will be extended in our further work.**

2) RECOGNITION ROBUSTNESS

For a practical system, robustness and reliability are two very important metrics. It is a pity that, similar to other research prototypes, our current implementation could not achieve self-adaption for different application scenarios, e.g., varying the distance between hands and sensing units, and users' diversity. However, it is feasible to improve its robustness with more advanced design in both hardware and software of learning algorithms. For the hardware of sensing units, we may employ more powerful Automatic Gain Control (AGC) amplifier to improve the detection sensitivity when the user's hand is a bit far away from the sensing

units so that we can obtain better sensing data to feed the recognition model. **At the software side, we can implement an incremental or online scheme to update the trained deep learning model whenever a new training sample is collected to automatically tune the model to adapt a new application scenario without the need of retraining the whole model.**

3) LIGHT SOURCE AND ENVIRONMENTAL REQUIREMENTS

As our system in fact senses the shadow variance caused by user's hand, the illumination and position of light source will certainly affect its performance. To let our current prototype work effectively, our current design need the light source stay upon the prototype and the illuminance level is a normal requirement of around 200 lux. If the light source is not upon our prototype, it is needed to properly position the sensing units to cover the shadow of hands and retrain the model. And we believe this is an affordable cost of using this novel technology to properly deploy and initialize the system, as a normal procedure to use a new equipment.

V. CONCLUSION

In this paper, we focus on the task of hand gesture recognition based on unmodulated ambient light by using RNN. We build a prototype consisting of an FPGA development board, an 8-channel ADC daughter board and 8 low cost photodiodes to capture the light intensity change cased by hand. **To infer the hand gesture from the sensed data, we propose an effective RNN-based learning architecture, in which three classical and representative RNN units are empirically selected. Through extensive field experiments under various settings, e.g., different sampling rates and training intensity, we find the deep learning model with GRU unit achieves the best tradeoff between accuracy and computational latency.** The overall recognition accuracy is beyond 99% with a low latency of 0.0622 s, strongly demonstrating its effectiveness and efficiency. **In order to put this novel technology into practical applications, we are also on the way to extend our research, on the one hand, to improve the robustness for self-adaption to various application scenarios, on the other hand, to boost the recognition ability for more types of hand gestures, e.g., combinational gesture with involving several fingers at the same time and dynamic hand gestures.**

ACKNOWLEDGMENT

The authors would like to thank JLC Co. for the support of producing circuit boards, and thank ALINX for the FPGA board utilized in the experiments.

REFERENCES

[1] T. Sharp, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, S. Izadi, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, and A. Vinnikov, "Accurate, robust, and flexible real-time hand tracking," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. (CHI)*, 2015, pp. 3633–3642.

[2] C. Zimmermann and T. Brox, "Learning to estimate 3D hand pose from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4913–4921.

[3] R. Wang, S. Paris, and J. Popović, "6D hands: Markerless hand-tracking for computer aided design," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2011, pp. 549–558.

[4] S. Gupta, D. Morris, S. Patel, and D. Tan, "SoundWave: Using the Doppler effect to sense gestures," in *Proc. ACM SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, New York, NY, USA, 2012, pp. 1911–1914, doi: 10.1145/2207676.2208331.

[5] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine–grained finger tracking," in *Proc. ACM CHI Conf. Hum. Factors Comput. Syst. (CHI)*, New York, NY, USA, 2016, pp. 1515–1525, doi: 10.1145/2858036.2858580.

[6] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. 19th Annu. Int. Conf. Mobile Comput., Netw. (MobiCom)*, 2013, pp. 27–38.

[7] M. Shahzad and S. Zhang, "Augmenting user identification with WiFi based gesture recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018, vol. 2, no. 3, pp. 134-1–134-27.

[8] L. Sun, S. Sen, D. Koutsonikolas, and K.-H. Kim, "WiDraw: Enabling hands-free drawing in the air on commodity WiFi devices," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2015, pp. 77–89.

[9] B. Kellogg, V. Talla, and S. Gollakota, "Bringing gesture recognition to all devices," in *Proc. 11th USENIX Conf. Netw. Syst. Design Implement. (NSDI)*, 2014, pp. 303–316.

[10] Z. Tian, J. Wang, X. Yang, and M. Zhou, "WiCatch: A Wi–Fi based hand gesture recognition system," *IEEE Access*, vol. 6, pp. 16911–16923, 2018.

[11] T. Li, X. Xiong, Y. Xie, G. Hito, X.-D. Yang, and X. Zhou, "Reconstructing hand poses using visible light," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Sep. 2017, vol. 1, no. 3, pp. 71-1–71-20, doi: 10.1145/3130937.

[12] C. Zhang, J. Tabor, J. Zhang, and X. Zhang, "Extending mobile interaction through near–field visible light sensing," in *Proc. ACM 21st Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, 2015, pp. 345–357, doi: 10.1145/2789168.2790115.

[13] A. Varshney, A. Soleiman, L. Mottola, and T. Voigt, "Battery-free visible light sensing," in *Proc. 4th ACM Workshop Visible Light Commun. Syst. (VLCS)*, New York, NY, USA, 2017, pp. 3–8, doi: 10.1145/3129881.3129890.

[14] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "WiFi-enabled device-free gesture recognition for smart home automation," in *Proc. IEEE 14th Int. Conf. Control Autom. (ICCA)*, Jun. 2018, pp. 476–481.

[15] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. 29th Annu. Symp. User Interface Softw. Technol. (UIST)*, 2016, pp. 851–860.

[16] C. Zhao, K.-Y. Chen, M. T. I. Aumi, S. Patel, and M. S. Reynolds, "Sideswipe: Detecting in-air gestures around mobile devices using actual GSM signal," in *Proc. 27th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2014, pp. 527–534.

[17] Y. Yang, J. Hao, J. Luo, and S. J. Pan, "CeilingSee: Device-free occupancy inference through lighting infrastructure based LED sensing," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2017, pp. 247–256.

[18] K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic Doppler sonar," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 1889–1892.

[19] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "AudioGest: Enabling fine-grained hand gesture detection by decoding echo signal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, New York, NY, USA, 2016, pp. 474–485, doi: 10.1145/2971648.2971736.

[20] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. ACM 22nd Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, 2016, pp. 82–94, doi: 10.1145/2973750.2973764.

[21] T. Li, C. An, Z. Tian, A. T. Campbell, and X. Zhou, "Human sensing using visible light communication," in *Proc. ACM 21st Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, 2015, pp. 331–344, doi: 10.1145/2789168.2790110.

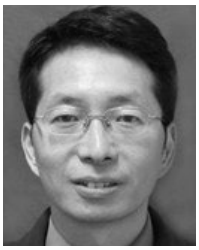
[22] V. Nguyen, M. Ibrahim, S. Rupavatharam, M. Jawahar, M. Gruteser, and R. Howard, "Eyelight: Light-and-shadow-based occupancy estimation and room activity recognition," in *Proc. IEEE Conf. Comput. Commun. INFOCOM*, Apr. 2018, pp. 351–359.
[23] D. Ma, G. Lan, M. Hassan, W. Hu, M. B. Upama, A. Uddin, and M. Youssef, "SolarGest: Ubiquitous and battery-free gesture recognition using solar cells," in *Proc. ACM 25th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, 2019, pp. 1–15, doi: 10.1145/3300061.3300129.
[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[25] Z. C. Lipton, "A critical review of recurrent neural networks for sequence learning," *CoRR*, vol. abs/1506.00019, Oct. 2015. [Online]. Available: http://arxiv.org/abs/1506.00019
[26] C. L. Giles, G. M. Kuhn, and R. J. Williams, "Dynamic recurrent neural networks: Theory and applications," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 153–156, Mar. 1994.
[27] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, Sep. 2014. [Online]. Available: https://arxiv.org/abs/1406.1078
[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[29] Y. Bengio, P. Frasconi, and P. Simard, "The problem of learning long-term dependencies in recurrent networks," in *Proc. IEEE Int. Conf. Neural Netw.*, Dec. 1993.
[30] S. El Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies," in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, pp. 493–499.
[31] G. Weiss, Y. Goldberg, and E. Yahav, "On the practical computational power of finite precision RNNS for language recognition," *CoRR*, vol. abs/1805.04908, May 2018. [Online]. Available: http://arxiv.org/abs/1805.04908
[32] F. Chollet, "Keras: The python deep learning library," *Astrophysics Source Code Library*, 2018.
[33] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

HAIHAN DUAN received the B.E. degree from East China Normal University, China. He is currently pursuing the M.E. degree with the College of Computer Science, Sichuan University, China. His research interests include smart sensing, computer vision, and deep learning.

MIAO HUANG received the B.E. degree from Sichuan Normal University, China. She is currently pursuing the M.E. degree with the College of Computer Science, Sichuan University, China. Her research interests include visible light communication and deep learning.

YANBING YANG received the B.E. and M.E. degrees from the University of Electronic Science and Technology of China, China, and the Ph.D. degree in computer science and engineering from Nanyang Technological University, Singapore. He is currently an Associate Research Professor with the College of Computer Science, Sichuan University, China. His research interests include the IoT, visible light communication, visible light sensing, and their applications.

JIE HAO received the B.S. degree from the Beijing University of Posts and Telecommunications, China, in 2007, and the Ph.D. degree from the University of Chinese Academy of Sciences, China, in 2014. From 2014 to 2015, she worked as a Postdoctoral Research Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore. She is currently an Assistant Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. Her research interests are wireless sensing and visible light communication.



LIANGYIN CHEN received the Ph.D. degree from the School of Computer Science, Sichuan University, in 2008. From 2009 to 2010, he was a Visiting Researcher with the University of Minnesota, under the supervision of Prof. T. He. He is currently a Professor with Sichuan University. He has authored or coauthored more than 40 articles, many of which were published in premier network journals and conferences. His research interests include wireless sensor networks, embedded systems, computer networks, distributed systems, big data analytics, natural language processing, the Internet of Things, and the industrial Internet.