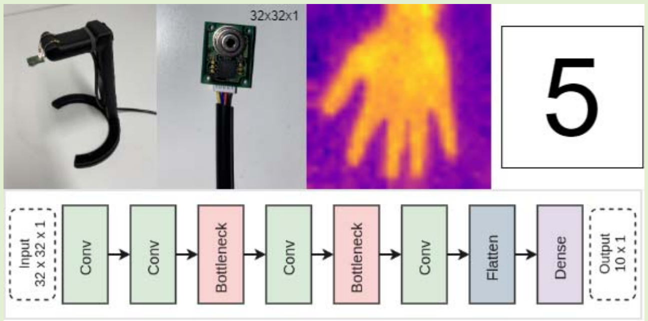
Deep Learning-Based Sign Language Digits Recognition From Thermal Images With Edge Computing System

Daniel S. Breland, Simen B. Skriubakken, Aveen Dayal, Ajit Jha,
Phaneendra K. Yalavarthy, *Senior Member, IEEE*,
and Linga Reddy Cenkeramaddi, *Senior Member, IEEE*



***Abstract*—The sign language digits based on hand gestures have been utilized in various applications such as human-computer interaction, robotics, health and medical systems, health assistive technologies, automotive user interfaces, crisis management and disaster relief, entertainment, and contactless communication in smart devices. The color and depth cameras are commonly deployed for hand gesture recognition, but the robust classification of hand gestures under varying illumination is still a challenging task. This work presents the design and deployment of a complete end-to-end edge computing system that can accurately provide the classification of hand gestures captured from thermal images. A thermal dataset of 3200 images was created with each sign language digit having 320 thermal images. The solution presented here utilizes live images taken from a low-resolution thermal camera of 32×32 pixels, feeding into a novel light weight deep learning model based on bottleneck motivated from deep residual learning for classification of hand gestures. The edge computing system presented here utilizes Raspberry pi with a thermal camera making it highly portable. The designed system achieves an accuracy of 99.52% on the test data set with an added advantage of accuracy being invariable to background lighting conditions as it is based on thermal imaging.**

***Index Terms*—Thermal imaging, gesture recognition, embedded systems, deep learning, neural networks, contactless applications, sign language digits.**

I. Introduction

The sign language based on hand gestures has applications in several areas such as automotive user interfaces, health and medical systems, health assistive technologies, crisis management and disaster relief, entertainment, and human-computer/robot interaction. This enables an effective contactless communication in many cases such as health care applications [1], [2], speaking, listening [3], entertainment such as gaming [4]–[6], contactless control of smart devices such as television [7], robotics [8], crisis management and disaster relief and several other areas including medical diagnostics and surveillance [9].

Recently, several deep learning models have been proposed for hand gesture recognition. A deep convolutional neural networks (CNN) based static hand gesture recognition was proposed in [10] using two publicly available datasets, National University of Singapore (NUS) hand posture dataset and American finger spelling A dataset. In this work, raw RGB images of hand postures have been utilized with CNN model providing the classification accuracy of 94.6%. A dynamic hand gesture recognition using two-dimensional (2D) CNN has been proposed in [11]. A combination of low and high resolution sub-networks have been deployed to obtain a classification accuracy of 98.2%. A CNN model along with

Manuscript received February 17, 2021; accepted February 19, 2021. Date of publication February 23, 2021; date of current version April 5, 2021. This work was supported by the Indo-Norwegian collaboration in Autonomous Cyber-Physical Systems (INCAPS) of International Partnerships for Excellent Education, Research and Innovation (INTPART) program from the Research Council of Norway under Project 287918. The associate editor coordinating the review of this article and approving it for publication was Dr. Sanket Goel. *(Corresponding author: Linga Reddy Cenkeramaddi.)*

Daniel S. Breland, Simen B. Skriubakken, Aveen Dayal, and Linga Reddy Cenkeramaddi are with the Department of Information and Communication Technology, University of Agder, 4879 Grimstad, Norway (e-mail: danisb15@student.uia.no, simebs16@student.uia.no, aveendayal97@gmail.com; linga.cenkeramaddi@uia.no).

Ajit Jha is with the Department of Engineering Science, University of Agder, 4879 Grimstad, Norway (e-mail: ajit.jha@uia.no).

Phaneendra K. Yalavarthy is with the Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru 560 012, India (e-mail: yalavarthy@iisc.ac.in).

Digital Object Identifier 10.1109/JSEN.2021.3061608

3D receptive fields for dynamic hand gesture recognition was proposed in [12]. Classification accuracy of 97.5% has been achieved with this hybrid neural network. A static hand gesture recognition using CNN with data augmentation has been proposed in [13]. Using this approach, an accuracy of 97.12% has been obtained. The same task using morphological filters was attempted in [14]. Classification accuracy of 96.83% has been achieved using this filtering approach. A dynamic hand gesture recognition has been attempted using the CNN and recurrent neural network (RNN) in [15]. With this approach, an overall accuracy of 85.46% was obtained. A hand gesture recognition using Gaussian mixture model was also attempted earlier in [16]. Using this model, an average recognition accuracy of 95.96% has been achieved. A deep CNN has been proposed for the task of hand gesture detection and recognition in [17]. Using this method, recognition accuracy of 90.7% was reported. A double channel (DC) CNN was also developed for hand gesture recognition in [18]. The recognition rate of 98.02% has been obtained in this approach. A three-dimensional (3D) CNN based hand gesture recognition was also developed and presented in [19]. It has achieved a classification accuracy of 77.5%. A multimodal hand gesture recognition using 3D CNNs was attempted in [20]. It achieved classification accuracy of 94.4%. A 3D CNN and long short-term memory (LSTM) based dynamic hand gesture recognition has been proposed in [21]. This approach has achieved an accuracy of 97.8%. A Tiny hand gesture recognition without localization using a deep CNN has been proposed in [22]. It achieved an accuracy of 97.1% with simple backgrounds and 85.3% for complex backgrounds. A hand gesture recognition using a compact CNN and surface electromyography signals has been proposed in [23]. With this compact CNN approach, an accuracy of 98.81% was reported. Image processing techniques along with a deep CNN has been proposed for hand gesture recognition in [24]. It achieved recognition rate of 95.61%. An online detection and classification of dynamic hand gestures with recurrent 3D CNNs was presented in [25]. Using this recurrent 3D CNNs, an accuracy 88.4% has been achieved. A probabilistic combination of CNN and RNN based hand gesture recognition resulted in an accuracy of 89.5% [26]. A Hand gesture recognition utilizing shape and texture evidences in complex background has been proposed in [27] with reported classification accuracy being 94.6%.

Majority of the developed models consist of several convolution layers, max-pooling layers, and different regularization layers like drop out, and L2 regularization. In addition, these models are heavy, which require more memory and greater model inference time. Memory efficient CNNs execution without having any compromise on the accuracy has been a challenge, especially when the inference has to be performed on an edge computing device in real time. The state-of-the-art performance has been achieved only with the help of neural network accelerators [28]. In this work, we develop a memory efficient CNN model which provides high accuracy without the need for accelerators. We also compare the performance of proposed memory efficient CNN model with the recently reported CNN models and the Big Transfer (BiT) model [29], [30], the state of the art model for image classification task on CIFAR10 dataset. This dataset is very similar to the developed thermal dataset with image size being (32,32,3). This work also shows that the proposed light-weight CNN out performs the state of the art model.

The above reported deep learning models were based on RGB images. The RGB cameras capture visible light and produce greyscale or RGB images, which can be utilized for many day-to-day applications including hand gesture recognition. Nevertheless, the quality of RGB images such as colors and visibility of the objects in the imaging scene largely depend on the background lighting conditions, such as artificial or natural light. Some attempts have been made for hand gesture recognition using low resolution RGB images under variable illumination conditions [31], [32] and depth images in low-intensity environments [33]. However, RGB cameras fail to capture any object in the imaging scene when it is totally dark. The illumination with changing intensity and color balance are some of the challenges associated with the quality of RGB images. Other sensors such as 3D sensors and near infrared sensors [34] have been introduced to overcome these limitations associated with RGB cameras. These sensors are active sensors with illumination being not a major concern. A passive sensor is preferred in many real-world applications. The micro electromechanical systems (MEMS) camera is a passive non-contact temperature measurement sensor. A MEMS thermal camera has two main parts, first one is the silicon lens and the second one is the thermopile sensor. The radiant heat of far-infrared rays emitted from objects in the imaging scene is focused on the thermopile sensor using a silicon lens [35]. The thermopile sensor produces electromotive force as per the incident radiant energy of far-infrared rays emitted from objects. Using the produced electromotive force and the internal thermal sensors, the temperature of objects in the imaging scene can be measured in a noncontact manner.

To the best of our knowledge, this is the first work to report thermal imaging based hand gestures, which are independent of background lighting including dark light conditions. We also propose a novel light weight deep learning model based on bottlenecks motivated from deep residual learning for classification of hand gestures using images obtained from a thermal camera.

The remainder of this article has been organized as follows. Section II provides the thermal imaging system description, and the experimental setup details used in order to obtain the hand gestures of the thermal images. Section III presents dataset details and variations associated with the data. Section IV elaborates on the machine learning model and dataset considered in this work. Section V presents the results. Finally, concluding remarks and future work was discussed in Section VI.

II. Thermal Imaging System Design

A MEMS thermal camera of Omron D6T has been utilized to create a dataset of thermal images for sign digits based on hand gestures. It has inter integrated circuit (I2C) interface and is connected to Raspberry Pi embedded system as shown in Fig. 1. To be able to interface the thermal camera to the

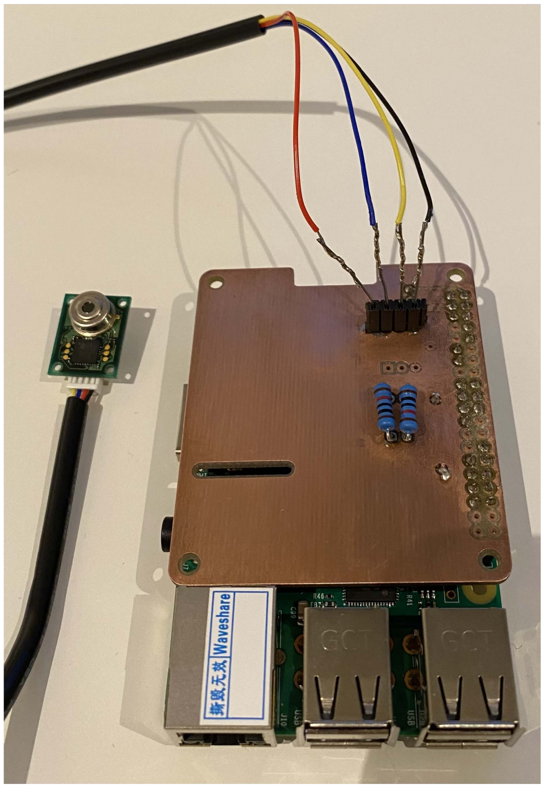


Fig. 1. Photograph showing the thermal camera coupling with the edge computing embedded system Raspberry Pi.

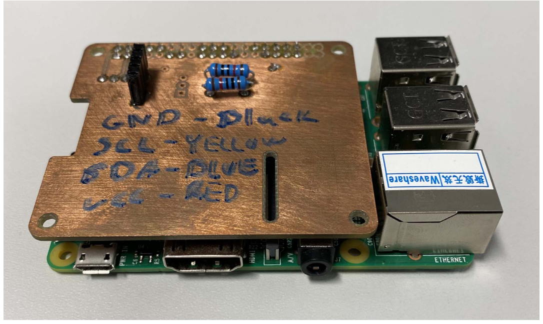


Fig. 2. Photograph showing the Raspberry Pi camera shield.

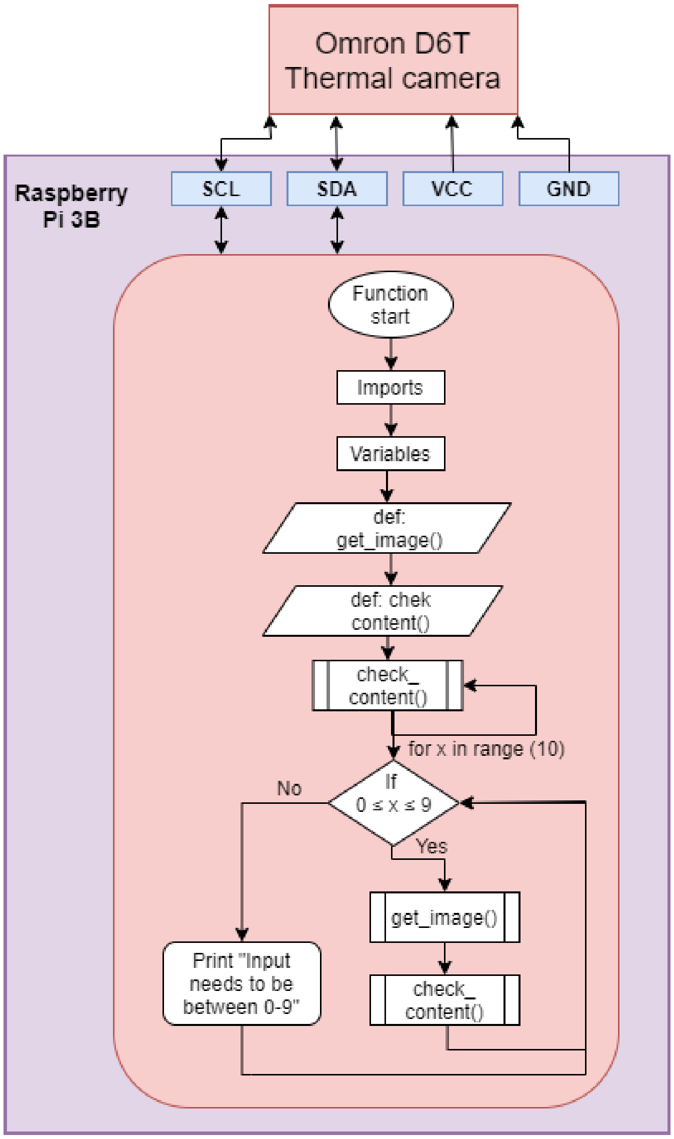


Fig. 3. Flowchart showing the important steps for capturing the thermal images with Raspberry Pi.

Raspberry Pi, it is necessary to design a custom shield. This shield is mounted to the pins of the Raspberry Pi with a pull up resistor for the thermal camera to function correctly. The pull up resistor also ensures the signal to and from the camera is correct. In addition, this blocks out any biased power or signals that may interfere. The resistors are connected to the power in one end, the data port and clock port on the other end. The wires are connected as show in Fig. 2. From top to bottom, they are connected to the ground, clock, data and power.

The Raspberry Pi shield is custom designed using Altium Designer [36] and a prototype has been made. The custom shield has only two layers, top and bottom. The custom shield also ensures that the wires are connected to the corresponding pins on the Raspberry Pi. For the proper operation of I2C interface, it needs to be connected to pin 3 (SDA) and pin 5 (SCL), also called GPIO 02 and GPIO 03 respectively. The power supply pins are pin 2 and pin 4, which supplies 5V, and it needs to be connected to one of these. The grounding has a lot of options, pin 6, pin 9, pin 14, pin 20, pin 25, pin 30, pin 34 and pin 39 and any one of these can be used to provide a common ground level [37], [38].

Once the Omron D6T thermal camera has been interfaced to the Raspberry pi, the thermal image dataset of hand gestures has been created by taking images from various person's hands and placed them in separate folders. The image folders are named from 0 to 9 and placed the images corresponding to the sign language digits. A custom software interface was developed to capture the thermal images using Raspberry Pi embedded system. A flowchart showing the complete steps for creating the dataset is shown in Fig. 3.

III. Thermal Imaging Dataset of Hand Gestures

Using the steps shown in Fig. 3, a large dataset of thermal images of hand gestures has been created. The complete dataset consists of 3200 images. For each sign digit, 320 hand

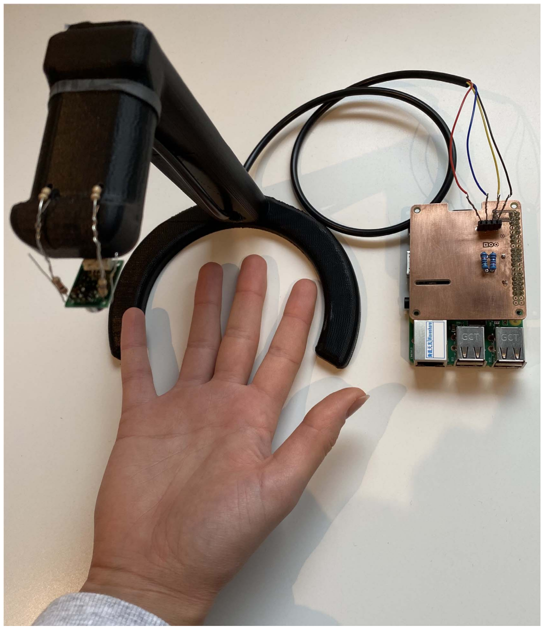


Fig. 4. Complete thermal imaging setup showing an example image collection.

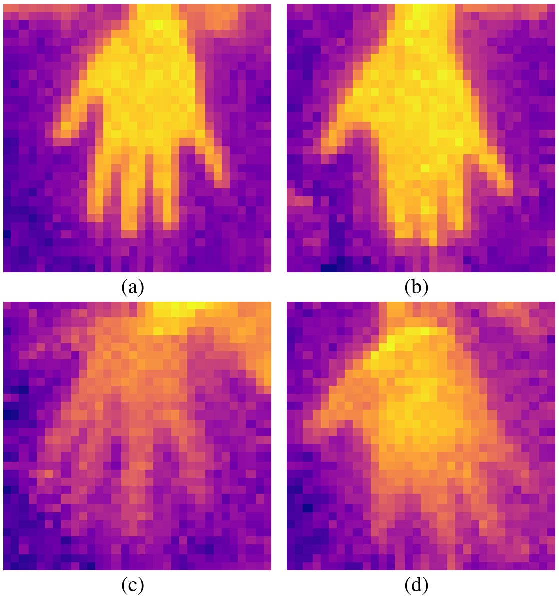


Fig. 5. Example thermal images in terms of variation in quality: (a) Good quality thermal image; (b) Good quality, bad positioning; (c) Poor quality, good positioning; and, (d) Varying quality from hand-palm to fingers.

gestures of thermal images were collected from different persons hands. While taking the data, hand is oriented in several ways with respect to the camera to be able to provide large variations to the dataset.

In order for the Omron D6T thermal camera to be placed in a stable position, it was attached to a tripod. The distance and FoV (Field of View) was optimized. The tripod was custom designed based on 3D-printing of a pre-drawn model [39]. The total finished setup can be seen in Fig. 4. The total setup is quite flexible and highly portable.

The Omron D6T thermal camera has a resolution of 32×32 pixels. This low resolution reduces the information to be provided for a classical machine learning, which requires hand-crafted features, making it challenging for the classification task. The total disc space of all 3200 images is only 8.2MB, a very low amount of data that can be easily migrated and utilized for any artificial intelligence model development. As it can be observed in Fig. 5, the quality of images may vary depending on the position and quality of the sensor. The image quality also depends on the external factors other than the thermal camera sensor itself. The thermal camera will detect difference in temperature from the background and the object in front, reflecting how warm the hand of the person is compared to the surface beneath. A hand gesture belonging to the same class taken from different persons is shown in Fig.5. The Fig. 5(a) shows a good quality image, where all fingers are visible and the hand is center placed. This will be the desired quality for images, although Fig. 5(b) gives a distinct color difference but, the fingers are not distinct due to the posture of the hand itself. In both Fig. 5(c) and 5(d), the contrast is poor due to non-uniform temperature on the hands itself. This poses several challenges for the classical machine learning algorithms to be able to classify irrespective of hands with different temperature and posture. For each person, the positioning of the hand is different, as hands geometry is distinct for each person along with gesture being different as well. This creates a large diversity in the dataset even for the same class and poses several challenges to provide robust classification with good accuracy. Therefore, it is important to have large number of images with large diversity in the data including capture of images from different persons.

Typical machine learning algorithm requires hand-crafted features to perform the classification task, in turn requiring good quality images. As there is large variation of temperature between different persons under same ambient conditions, the quality of thermal images will also vary. A high quality camera increases the stability, thus removing the variations in image quality with added advantage of improving the speed of thermal image acquisition. The thermal images that were utilized in this work are of very low resolution, which makes the classification extremely challenging. The details in thermal images can be completely missed with only having separate color for hand and background. The images in Fig. 6 shows a complete set of thermal images, that has been taken with Omron D6T thermal camera. With such a low resolution camera, the temperature difference needs to be higher for the camera to separate objects from the background. In all images in Fig. 6, there is some leakage of heat, as the camera might pick up some heat from the table. This will manifest as purple color in the thermal images.

Some images were also alike due to this low resolution, example being thermal images of number 3, 6 and 9, corresponding to image 6(d), 6(g) and 6(j) in Fig. 6. For a trained personnel, this is an easy classification task. A machine (computer) on the other hand, needs efficient algorithm to be able to classify correctly. A set of good quality images will

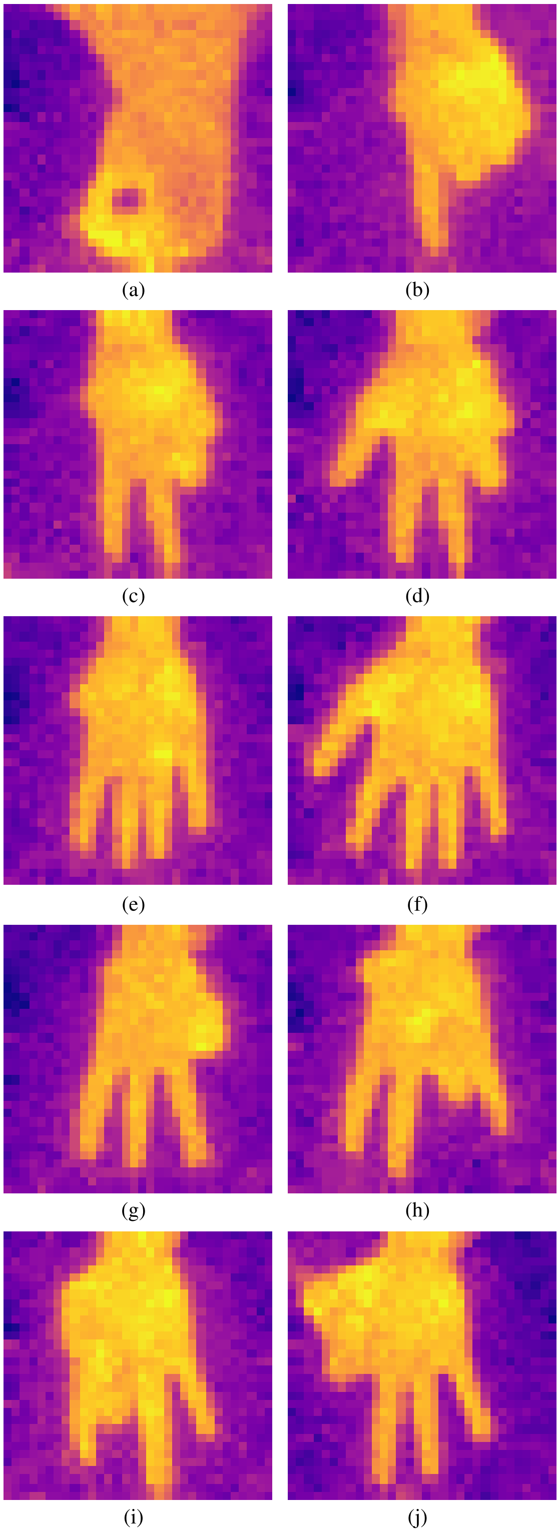


Fig. 6. A complete set of thermal images: (a) Thermal image, number 0; (b) Thermal image, number 1; (c) Thermal image, number 2; (d) Thermal image, number 3; (e) Thermal image, number 4; (f) Thermal image, number 5; (g) Thermal image, number 6; (h) Thermal image, number 7; (i) Thermal image, number 8; and, (j) Thermal image, number 9.

increase the success of detecting the correct gesture, for both humans and machine.

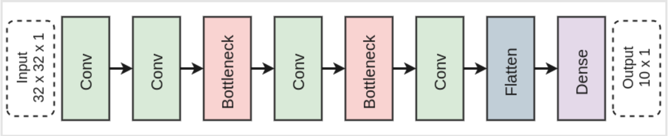


Fig. 7. Architecture of the proposed light weight CNN model. All ‘Conv’ layers use a kernel size of ‘3 × 3’ followed by batch normalization operation and ‘ReLU’ activation function. The details of layers were presented in Table II.

TABLE I
NUMBER OF SAMPLE IN TRAINING, VALIDATION AND TEST DATASET FOR EACH CLASS OF THE DEVELOPED THERMAL IMAGING DATASET

Class	Training	Validation	Test
0	231	25	64
1	231	25	64
2	231	25	64
3	231	25	64
4	231	25	64
5	231	25	64
6	231	25	64
7	231	25	64
8	231	25	64
9	231	25	64

After collecting these 3200 images (320 samples for each of the ten classes), we divided this dataset into three parts, training, validation and testing part. The testing dataset consists of 20% of the entire dataset such that each class has equal number of samples. From the remaining dataset, we randomly sample 10% to create a validation dataset. The number of samples in each of the dataset for each class has been shown in the Table I.

IV. LIGHT-WEIGHT DEEP LEARNING MODEL FOR HAND GESTURE RECOGNITION

A. Model

As discussed earlier, the classical machine learning model requires hand-crafted features and are not robust in performing classification task with low-quality thermal images. The superior alternative is to develop a deep learning model, which is fully data-driven and learns the feature set based on the training dataset. The proposed deep learning model is extremely light-weight and efficient with number of parameters being 851,978 and having a size of 10 Megabytes. The model consists of several bottleneck layers, which were inspired from deep residual networks (architecture was provided in Fig. 7 and layers details were presented in Table II). Each bottleneck layer was composed of stacked residual blocks as shown in Fig. 8 with layers details being provided in Table III. The first block performs three operations, namely depth squeezing using a ’1 × 1’ convolution, local feature extraction using ’3×3’ convolution and finally depth stretching using ’1 × 1’ convolution. The second block also performs 3 similar operations with the only difference of using dilated convolution of size ’3 × 3’ to extract local features instead of regular convolution. This dilation increases the receptive field of the convolution network. In short, the operations performed in the proposed bottleneck can be summarized as follows: Given a feature map x (as the input of bottleneck), the output

TABLE II
DETAILS OF DIFFERENT LAYERS IN THE PROPOSED LIGHTWEIGHT CNN (ARCHITECTURE WAS SHOWN IN FIG. 7. EACH CONV LAYERS REPRESENTS 3 × 3 CONVOLUTION, FOLLOWED BY RELU ACTIVATION ALONG WITH BATCH NORMALIZATION. THE BOTTLENECK BLOCK WAS PRESENTED IN FIG. AND DETAILS WERE PROVIDED IN TABLE III

Layer Type	Output Size
Input	(None, 32, 32, 1)
Conv	(None, 15, 15, 64)
Conv	(None, 7, 7, 128)
Bottleneck	(None, 7, 7, 128)
Conv	(None, 3, 3, 256)
Bottleneck	(None, 3, 3, 256)
Conv	(None, 2, 2, 128)
Flatten	(None, 512)
Dense	(None, 10)

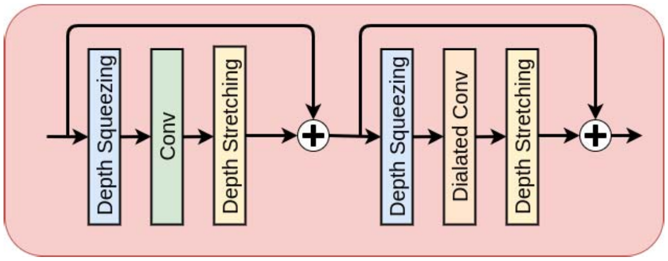


Fig. 8. Architecture of the bottleneck module. Details of each layer was provided in Table III.

TABLE III
DIFFERENT LAYERS IN THE BOTTLENECK BLOCK UTILIZED IN THE PROPOSED MODEL MENTIONED IN TABLE II. 'M' IS SPATIAL EXTENT AND 'Z' IS DEPTH OF THE FEATURE MAPS. THE ARCHITECTURE OF THE BOTTLE NECK WAS PROVIDED IN FIG. 8

Layer Type	Output Size
Input	(None, M, M, Z)
1x1 Conv	(None, M, M, Z/4)
3x3 Conv	(None, M, M, Z/4)
1x1 Conv	(None, M, M, Z)
Add	(None, M, M, Z)
1x1 Conv	(None, M, M, Z/4)
3x3 Dilated Conv	(None, M, M, Z/4)
1x1 Conv	(None, M, M, Z)
Add	(None, M, M, Z)
Output	(None, M, M, Z)

$h(x)$ of the initial residual block can be written as:

$$h(x) = f(x, \theta_1) + x \tag{1}$$

where, $f(x, \theta_1)$ is sequence of convolution operations parameterized by θ_1, performing depth squeezing, feature extraction and depth stretching. Note that it is easy to optimize $f(x, \theta_1)$, than to learn the underlying $h(x)$ directly from x [40]. Finally, the output $b(x)$ of the bottleneck is given by:

$$b(x) = g(h(x), \theta_2) + h(x) \tag{2}$$

where, $g(x, \theta_2)$ is sequence of convolution operations parameterized by θ_2, performing depth squeezing, feature extraction via dilated convolution and depth stretching.

As seen in Table II, the bottleneck layers were placed alternatively along with other convolution layers. Each convolution step utilized in the entire network was followed by a RELU activation and a batch normalization step. The input images that were acquired from Raspberry pi-device connected with a thermal camera are of the size 32 × 32. Given a mini-batch with N samples, the cross-entropy loss $\mathcal{L}$ was computed as shown below:

$$\mathcal{L} = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\sum_{t=0}^{C-1} y_{it}\log(\hat{y_{it}}) \tag{3}$$

where y is the one hot encoded label ($C \times 1$) and $\hat{y}$ is the predicted softmax probabilities ($C \times 1$) and, C is number of classes.The proposed model is extremely light weight and can be easily trained from scratch instead of transfer learning and was successfully trained from the well known kaiming initialization [41]. The proposed model was also trained using Adam optimizer with a learning rate of '0.005' and with a batch size of 8. The end to end training of the model was performed on Google Colab using Keras deep learning library [42] with Tesla T4 GPU consisting of 14 GB GPU memory.

B. Benchmark Model

We compared our proposed model with Big Transfer (BiT) model, the state of the art model for image classification task on CIFAR10 dataset [29]. The CIFAR10 dataset is very similar to the thermal dataset proposed as each image in CIFAR10 has the size (32,32,3). We train the model using transfer learning technique, such as Fine Tuning. In this method, the model was first pre-trained on large dataset and then fine tuned to new dataset by freezing few initial layers of the model and training only the other unfrozen layers.

The Big transfer model was built using two components, namely upstream pre-training and downstream fine tuning. In the upstream pre-training the model was pre-trained on large datasets and in downstream fine tuning task, the pre-trained model was trained on new datasets using transfer learning. This combination of two components help the BiT model perform effective hyper parameter search and achieve state of art results on many downstream tasks [29], [30]. There are many variants in the BiT model, differing in their architecture size and the pre-trained dataset. Since our model should be deployed on an embedded device (raspberry pi), we use only the relatively light weight versions of the BiT model. We use the BiT-S models, pre-trained on ILSVRC-2012 dataset, with 50 × 3 and 101 × 3 resnet architectures. We also use BiT-M model, pre-trained on ImageNet-21K dataset, with 50 × 3 and 101 × 3 resnet architectures [30]. The details of these models in terms of number of parameters as model size are given in the last two columns of Table IV.

V. RESULTS AND DISCUSSION

The proposed light-weight CNN model achieves an accuracy of 99.52% on the test dataset. The confusion matrix of the proposed CNN model on the test dataset is shown in the Fig. 9. It can be observed from Fig.9 that the digits '1', '2', '3', '6', '7', '8' and '9' were able to correctly identify 100% of their respective class's test samples. Digits '0', '4' and '5'

TABLE IV
PERFORMANCE COMPARISON OF PROPOSED MODEL AND BENCHMARK MODELS IN TERMS OF TEST ACCURACY. THE MODEL SIZE IN MB (MEGA BYTES) AND NUMBER OF PARAMETERS ARE PROVIDED FOR EASY REFERENCE IN THE LAST TWO COLUMNS

S.No	Model	Accuracy	Model Size (MB)	Parameters
1	Proposed Model	99.52%	10	851,978
2	BiT-S(50x3)	75.16%	806	211,235,530
3	BiT-S(101x3)	70%	1000	381,851,338
4	BiT-M(50x3)	77.6%	806	211,235,530
5	BiT-M(101x3)	71.1%	1000	381,851,338

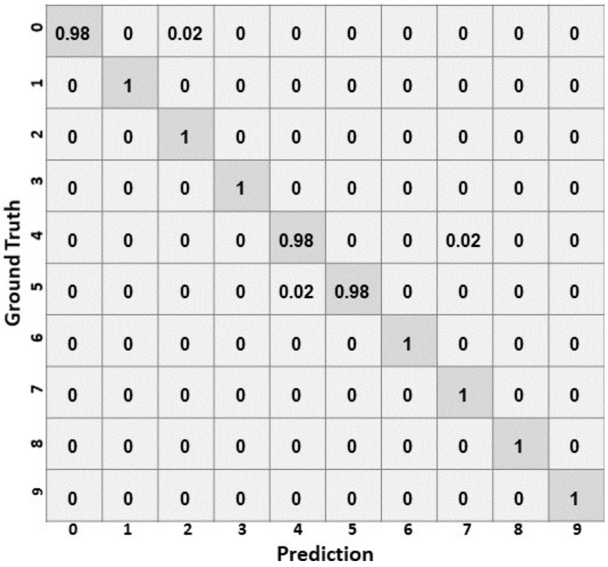


Fig. 9. Confusion Matrix of the proposed light weight CNN model for sign language digits (0 to 9) recognition from thermal images. The overall accuracy was provided in Table IV.

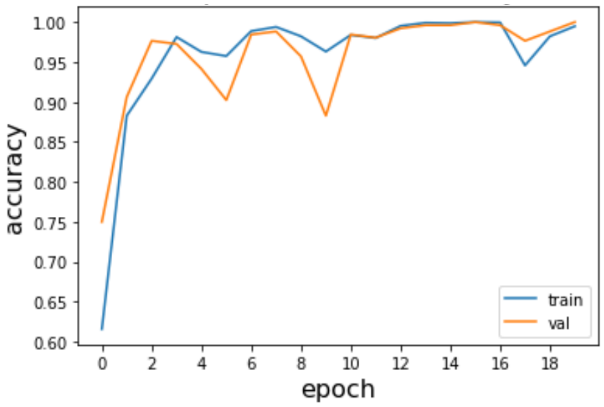


Fig. 10. Plot of Training and Validation accuracy of the proposed light weight CNN model as a function of epochs.

were able to correctly identify 98% of their respective class's test samples.

The training and validation accuracy plot of the proposed CNN model is also shown in the Fig. 10. As shown in the Fig. 10, the proposed model converges after 20 epochs to validation accuracy of 99.58%. The average inference time for the proposed model on one test sample is 30 mill sec on the Raspberry Pi edge computing device.

TABLE V
PERFORMANCE COMPARISON OF PROPOSED MODEL AND OTHER DEEP LEARNING MODELS IN TERMS OF TEST ACCURACY. EXCEPT FIRST 5 MODELS, REST MODELS HAVE UTILIZED ONLY RGB IMAGES

S.No	Model	Accuracy	Ref.
1	Proposed Model	99.52%	This work
2	BiT-S(50x3)	75.16%	BiT-S(50x3) [29], [30]
3	BiT-S(101x3)	70%	BiT-S(101x3) [29], [30]
4	BiT-M(50x3)	77.6%	BiT-M(50x3) [29], [30]
5	BiT-M(101x3)	71.1%	BiT-M(101x3) [29], [30]
6	Deep CNN	94.6%	[10]
7	2D CNN	98.2%	[11]
8	CNN with 3D Receptive fields	97.5%	[12]
9	CNN w/data augmentation	97. 2%	[13]
10	CNN w/morphological filters	96.83%	[14]
11	CNN and RNN	85.46%	[15]
12	Gaussian Mixture	95.96%	[16]
13	Deep CNN	90.7%	[17]
14	DC CNN	98.02%	[18]
15	3D CNN	77.5%	[19]
16	3D CNN	94.4%	[20]
16	3D CNN and LSTM	97.8%	[21]
17	Deep CNN	97.1%	[22]
18	Compact CNN	98.81%	[23]
19	Deep CNN and Image processing	95.61%	[24]
20	Recurrent 3D CNN	88.4%	[25]
21	CNN and RNN	89.5%	[26]

Comparing the performance of the proposed CNN model with the benchmark model, shown in the Table IV, it can be observed that the proposed model is very lightweight and also has very high accuracy. Performance comparison of proposed model and published CNN models in terms of test accuracy was shown in the Table V. it is evident from these results that the proposed model is quite accurate for the classification of hand gestures of thermal images and well suited for performing inference on the edge computing device.

As thermal cameras are becoming embedded part of mobile phones or being available as standard accessory [43], it is important that the developed models especially as assistive technologies can work on an edge device (including mobile phone). The light-weight CNNs have this distinct advantage being easily deployable in edge device and can be converted into mobile applications that can provide inference on a mobile phone. They have found applications in other domains, such as medical image processing [44] as well as other contactless operations [45]. From Table IV, it is evident that the proposed light-weight CNN has provided improved accuracy with at least three orders of magnitude less parameters, making it easily deployable in a mobile platform [44]. The improved accuracy can be attributed to bottlenecks utilized in the proposed model. These bottlenecks not only make the network light weight, it forces the network to compress feature representations to best fit in the available space, in order to provide improved training. This has been shown to provide better generalization on test data (unseen images) compared to standard heavy models [44], [45]. The same has been confirmed in our study as well.

The current study utilized very low quality thermal images (size of 32 × 32) with recognition being performed using a light-weight CNN model. Study involving varying lighting conditions was performed here. Study involving complex backgrounds was not attempted here as the main aim of this work was to show end-to-end thermal imaging based hand gesture recognition. Earlier studies involving RGB cameras have shown that utilization of hand-crafted features, such as shape and texture extracted from the gestures, were beneficial in providing improved accuracy with complex backgrounds [27]. A similar study will be taken up as a future work to provide a robust solution towards hand gestures recognition using thermal camera.

VI. Conclusion

This work presented the design of a complete end-to-end embedded system, which can accurately recognize the hand gestures of the low-resolution thermal images of 32×32 pixels. A thermal dataset of 3200 images was curated and each sign language digit has 320 hand gestures of thermal images. We have also developed a lightweight convolutional neural network to provide high accuracy and the need for having high performance computing environment. The designed system has achieved an accuracy of 99.52% on the test dataset with an added advantage of accuracy being invariable to background lighting conditions as it is based on thermal imaging. The developed system has shown that thermal imaging is well suited for the hand gesture recognition in dark light conditions.

References

[1] Z. N. Ahmed and J. A. Hussien, “An interactive and predictive pre-diagnostic model for healthcare based on data provenance,” *UHD J. Sci. Technol.*, vol. 3, no. 2, p. 59, Oct. 2019.

[2] O. Asan and E. Montague, “An interactive and predictive pre-diagnostic model for healthcare based on data provenance,” *Behav. Inf. Technol.*, vol. 33, no. 3, pp. 259–270, 2014.

[3] H. Kaur and J. Rani, “A review: Study of various techniques of hand gesture recognition,” in *Proc. IEEE 1st Int. Conf. Power Electron., Intell. Control Energy Syst. (ICPEICES)*, Jul. 2016, pp. 1–5.

[4] H. Kang, C. W. Lee, and K. Jung, “Recognition-based gesture spotting in video games,” *Pattern Recognit. Lett.*, vol. 25, no. 15, pp. 1701–1714, Nov. 2004.

[5] X. Zhang, X. Chen, W.-H. Wang, J.-H. Yang, V. Lantz, and K.-Q. Wang, “Hand gesture recognition and virtual game control based on 3D accelerometer and EMG sensors,” in *Proc. 14th Int. Conf. Intell. User Interface*, Feb. 2009, pp. 401–406.

[6] Y. Li *et al.*, “Hand gesture recognition and real-time game control based on a wearable band with 6-axis sensors,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6.

[7] D. K. Vishwakarma and R. Kapoor, “An efficient interpretation of hand gestures to control smart interactive television,” *Int. J. Comput. Vis. Robot.*, vol. 7, no. 4, pp. 454–471, 2017.

[8] Q. Lei, H. Zhang, Z. Xia, Y. Yang, Y. He, and S. Liu, “Applications of hand gestures recognition in industrial robots: A review,” in *Proc. 11th Int. Conf. Mach. Vis. (ICMV)*, vol. 11041, A. Verikas, D. P. Nikolaev, P. Radeva, and J. Zhou, Eds. Bellingham, WA, USA: SPIE, 2019, pp. 455–465, doi: 10.1117/12.2522962.

[9] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, “Vision-based hand-gesture applications,” *Commun. ACM*, vol. 54, no. 2, pp. 60–71, Feb. 2011, doi: 10.1145/1897816.1897838.

[10] V. Adithya and R. Rajesh, “A deep convolutional neural network approach for static hand gesture recognition,” *Procedia Comput. Sci.*, vol. 171, pp. 2353–2361, Jan. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050920312473

[11] F. Zhan, “Hand gesture recognition with convolution neural networks,” in *Proc. IEEE 20th Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Jul. 2019, pp. 295–298.

[12] H.-J. Kim, J. S. Lee, and J.-H. Park, “Dynamic hand gesture recognition using a CNN model with 3D receptive fields,” in *Proc. Int. Conf. Neural Netw. Signal Process.*, Jun. 2008, pp. 14–19.

[13] M. Z. Islam, M. S. Hossain, R. U. Islam, and K. Andersson, “Static hand gesture recognition using convolutional neural network with data augmentation,” in *Proc. Joint 8th Int. Conf. Informat., Electron. Vis. (ICIEV), 3rd Int. Conf. Imag., Vis. Pattern Recognit. (icIVPR)*, May 2019, pp. 324–329.

[14] R. F. Pinto, C. D. B. Borges, A. M. A. Almeida, and I. C. Paula, “Static hand gesture recognition based on convolutional neural networks,” *J. Electr. Comput. Eng.*, vol. 2019, Oct. 2019, Art. no. 4167890, doi: 10.1155/2019/4167890.

[15] K. Lai and S. N. Yanushkevich, “CNN+RNN depth and skeleton based dynamic hand gesture recognition,” in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3451–3456.

[16] H.-I. Lin, M.-H. Hsu, and W.-K. Chen, “Human hand gesture recognition using a convolution neural network,” in *Proc. IEEE Int. Conf. Automat. Sci. Eng. (CASE)*, Aug. 2014, pp. 1038–1043.

[17] P. S. Neethu, R. Suguna, and D. Sathish, “An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks,” *Soft Comput.*, vol. 24, no. 20, pp. 15239–15248, Oct. 2020, doi: 10.1007/s00500-020-04860-5.

[18] X. Y. Wu, “A hand gesture recognition algorithm based on DC-CNN,” *Multimedia Tools Appl.*, vol. 79, nos. 13–14, pp. 9193–9205, Apr. 2020, doi: 10.1007/s11042-019-7193-4.

[19] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, “Hand gesture recognition with 3D convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 1–7.

[20] N. N. Hoang, G.-S. Lee, S.-H. Kim, and H.-J. Yang, “A real-time multimodal hand gesture recognition via 3D convolutional neural network and key frame extraction,” in *Proc. Int. Conf. Mach. Learn. Mach. Intell. (MLMI)*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 32–37, doi: 10.1145/3278312.3278314.

[21] N. L. Hakim, T. K. Shih, S. P. K. Arachchi, W. Aditya, Y.-C. Chen, and C.-Y. Lin, “Dynamic hand gesture recognition using 3DCNN and LSTM with FSM context-aware model,” *Sensors*, vol. 19, no. 24, p. 5429, Dec. 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31835404

[22] P. Bao, A. I. Maqueda, C. R. del-Blanco, and N. García, “Tiny hand gesture recognition without localization via a deep convolutional network,” *IEEE Trans. Consum. Electron.*, vol. 63, no. 3, pp. 251–257, Aug. 2017.

[23] L. Chen, J. Fu, Y. Wu, H. Li, and B. Zheng, “Hand gesture recognition using compact CNN via surface electromyography signals,” *Sensors*, vol. 20, no. 3, p. 672, Jan. 2020, doi: 10.3390/s20030672.

[24] Y.-L. Chung, H.-Y. Chung, and W.-F. Tsai, “Hand gesture recognition via image processing techniques and deep CNN,” *J. Intell. Fuzzy Syst.*, vol. 39, no. 3, pp. 4405–4418, Oct. 2020, doi: 10.3233/JIFS-200385.

[25] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, “Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4207–4215.

[26] A. Tewari, B. Taetz, F. Grandidier, and D. Stricker, “[POSTER] a probabilistic combination of CNN and RNN estimates for hand gesture based interaction in car,” in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR-Adjunct)*, Oct. 2017, pp. 1–6.

[27] D. K. Vishwakarma, “Hand gesture recognition using shape and texture evidences in complex background,” in *Proc. Int. Conf. Inventive Comput. Informat. (ICICI)*, Nov. 2017, pp. 278–283.

[28] B. L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proc. IEEE*, vol. 108, no. 4, pp. 485–532, Apr. 2020.

[29] A. Kolesnikov *et al.*, "Big transfer (bit): General visual representation learning," in *Computer Vision*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 491–507.

[30] A. Kolesnikov *et al.*, "Large scale learning of general visual representations for transfer," *CoRR*, pp. 1–28, 2019. [Online]. Available: http://arxiv.org/abs/1912.11370

[31] D. K. Vishwakarma, R. Maheshwari, and R. Kapoor, "An efficient approach for the recognition of hand gestures from very low resolution images," in *Proc. 5th Int. Conf. Commun. Syst. Netw. Technol.*, Apr. 2015, pp. 467–471.

[32] D. K. Vishwakarma and R. Kapoor, "Simple and intelligent system to recognize the expression of speech-disabled person," in *Proc. 4th Int. Conf. Intell. Hum. Comput. Interact. (IHCI)*, Dec. 2012, pp. 1–6.

[33] D. K. Vishwakarma and V. Grover, "Hand gesture recognition in low-intensity environment using depth images," in *Proc. Int. Conf. Intell. Sustain. Syst. (ICISS)*, Dec. 2017, pp. 429–433.

[34] S. Chen, Y. Li, J. Zhang, and W. Wang, "Active vision sensors," in *Active Sensor Planning for Multiview Vision Tasks*. Springer, 2008, pp. 11–38. [Online]. Available: https://www.springer.com/gp/book/9783540770718

[35] L. R. Cenkeramaddi, J. Bhatia, A. Jha, S. K. Vishkarma, and J. Soumya, "A survey on sensors for autonomous systems," in *Proc. 15th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Nov. 2020, pp. 1182–1187.

[36] Altium Limited. (Nov. 3, 2020). *The Most Connected Experience for PCB Design and Realization*. [Online]. Available: https://www.altium.com/

[37] J. Marot and S. Bourennane, "Raspberry pi for image processing education," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 2364–2366.

[38] S. Mischie, "On teaching raspberry pi for undergraduate university programmes," in *Proc. 12th IEEE Int. Symp. Electron. Telecommun. (ISETC)*, Oct. 2016, pp. 149–153.

[39] THE_CRAFT_DUDE. (Oct. 26, 2020). *Simple Headphone Stand*. [Online]. Available: https://cults3d.com/en/3d-model/gadget/simple-headphone-stand

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Washington, DC, USA, Dec. 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.

[42] F. Chollet *et al.*, "Keras: The Python deep learning library," in *Proc. ASCL*, 2018, p. 1806.

[43] E. Villa, N. Arteaga-Marrero, and J. Ruiz-Alzola, "Performance assessment of low-cost thermal cameras for medical applications," *Sensors*, vol. 20, no. 5, p. 1321, Feb. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/5/1321

[44] N. Paluru *et al.*, "Anam-Net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 5, 2021, doi: 10.1109/TNNLS.2021.3054746.

[45] A. Dayal, N. Paluru, L. R. Cenkeramaddi, J. Soumya, and P. K. Yalavarthy, "Design and implementation of deep learning based contactless authentication system using hand gestures," *Electronics*, vol. 10, no. 2, p. 182, Jan. 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/2/182

Daniel S. Breland was born in Kristiansand, Norway, in 1996. He received the bachelor's degree in electronics engineering from the University of Agder, Grimstad, Norway, in 2019, where he is currently pursuing the master's degree in communication technologies with the Department of ICT. His main research interests include wireless communications and sensor networks in IoT environments.

Simen B. Skriubakken received the bachelor's degree in electronics from the University of Agder, in 2019, where he is currently pursuing the master's degree in communication technologies. His main research interests include sensor technology, wireless communications, and IoT solutions.

Aveen Dayal received the bachelor's degree in computer science and engineering in 2020. He currently works as a Visiting Research Student with the Department of ICT, University of Agder, Grimstad, Norway. His main research interest includes machine learning for autonomous cyber-physical systems.

Ajit Jha was born in Nepal, in 1984. He received the B.Sc. degree in electronics and communication engineering, Bangladesh, in 2007, and the European Master's degree in photonic networks from Aston University, U.K., and Scuola Superiore Sant Anna, Italy, and the Ph.D. degree from the Technical University of Catalunya, Spain, and the Karlsruhe Institute of Technology, Germany, in 2012 and 2016, respectively. From 2016 to 2019, he worked at various industries related to autonomous vehicle working on innovative technologies, such as automotive Ethernet, ADAS, surround view systems, camera mirror system, blind sport warning to name a few. He is currently an Associate Professor of Mechatronics with the Department of Engineering Sciences, University of Agder, Norway. He is actively involved in research focused on sensors, sensor fusion, image/signal processing, ML, ADAS functionalities towards autonomous systems, and the IoT. He has coauthored more than 20 articles and two patents. He has been an Active Reviewer, and a member of technical program committee of numerous international peer-reviewed journals and conferences. He was a recipient of Erasmus Mundus Masters Course (EMMC) and Erasmus Mundus Joint Doctorate (EMJD) both funded by European Union (EU).

Phaneendra K. Yalavarthy (Senior Member, IEEE) received the M.Sc. degree in engineering from the Indian Institute of Science, Bengaluru, India, and the Ph.D. degree in biomedical computation from the Dartmouth College, Hanover, NH, USA, in 2007. He is currently an Associate Professor with the Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru. His research interests include medical image computing, medical image analysis, and biomedical optics. He is a Senior Member of SPIE and OSA. He serves as an Associate Editor for the IEEE TRANSACTIONS ON MEDICAL IMAGING.

Linga Reddy Cenkeramaddi (Senior Member, IEEE) received the master's degree in electrical engineering from the Indian Institute of Technology, New Delhi, India, in 2004, and the Ph.D. degree in electrical engineering from the Norwegian University of Science and Technology, Trondheim, Norway, in 2011. He worked with Texas Instruments in mixed signal circuit design before joining the Ph.D. program at NTNU. From 2010 to 2012, he worked in radiation imaging for an atmosphere space interaction monitor (ASIM mission to International Space Station) with the University of Bergen, Norway. He is currently working as an Associate Professor with the University of Agder, Grimstad, Norway. His research interests include cyber-physical systems, autonomous systems, and wireless embedded systems.