Sign Language Gesture Recognition using Doppler Radar and Deep Learning

Hovannes Kulhandjian[†], Prakshi Sharma[†], Michel Kulhandjian[‡], Claude D'Amours[‡]

[†]Department of Electrical and Computer Engineering, California State University, Fresno, Fresno, CA 93740, U.S.A.
E-mail: hkulhandjian@csufresno.edu,prakshi1993@mail.fresnostate.edu
[‡]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada
E-mail: mkk6@buffalo.edu,cdamours@uottawa.ca

***Abstract*—In this paper, we study American sign language (ASL) hand gesture recognition using Doppler radar. A set of ASL hand gesture motions are captured as micro-Doppler signals using a microwave X-band Doppler radar transceiver. We apply joint time-frequency analysis and observe the presence of the micro-Doppler signatures in the spectrogram. The micro-Doppler signatures of different hand gestures are analyzed using Matlab. Each hand gesture is observed to contain unique spectral characteristics. Based on unique spectral characteristics, we investigate the classification of ASL essential short phrases including emergency signals. For recognizing and characterizing the presence of micro-Doppler signatures in spectrogram we explore deep convolution neural network (DCNN) algorithm. We show that the DCNN algorithm can classify different sign language gestures based on the presence of micro-Doppler signatures in the spectrogram with fairly high accuracy. Experimental results reveal that utilizing 80% of data for training, and the remaining 20% for validation purposes in DCNN algorithm a validation accuracy of 87.5% is achieved. To further improve the recognition system, we apply a very deep learning algorithm VGG-16 using transfer learning, which improves the validation accuracy to 95%.**

***Index Terms*—Detection and classification, American sign language (ASL) gesture recognition, Doppler radar, micro-Doppler signatures, deep convolution neural network (DCNN), VGG-16 algorithm.**

I. Introduction

Hand gesture recognition has many applications ranging from medical, gaming, human machine interaction as well as sign language interpretation [1]–[3]. The problem of hand gesture recognition consists of identifying a given gesture performed by hand movements. There are various ways that can be used to perform hand gesture recognition ranging from video or image processing to radar motion detection and tracking [4]. A number of research works have studied sign language hand gesture recognition using video or image signal processing with the combination of machine learning.

In [5], radar is used to enable gesture recognition based on the micro-Doppler signatures that are associated to different movements. Five micro-Doppler based handcrafted features are used for gesture recognition. A simple k-nearest neighbor (kNN) classifier [6] is applied to evaluate the importance of the five features. The overall classification accuracy of the proposed framework was 84%.

In [7], a method is presented to classify four different kinds of hand gestures that include snapping fingers, flipping fingers, hand rotation and calling, using a radar micro-Doppler sensor. Two different kinds of micro-Doppler features are extracted from time-frequency spectrum and support vector machine (SVM) [6] is applied to classify the four kinds of gestures. Experimental results reveal the proposed method classification accuracy was 88.6%.

In [8], deep neural network is applied for American sign language (ASL) fingerspelling (posture) translation purposes. The 'Kaggle' ASL letter database of hand gestures was used to evaluate the framework. Performance validation provides high accuracy posture translation.

In [9], a real-time ASL fingerspelling translator based on convolutional neural network (CNN) is presented. A model is developed for classification of letters from $a-e$ correctly with first-time users and another that classifies letters from $a-k$ correctly in the majority of cases.

In [10], hand gesture recognition using radar micro-Doppler signature envelopes is presented. The kNN classifier and Manhattan distance (ℓ_1) [11] measure is used in their algorithm for distinguishing the envelope values. The algorithm uses an energy-based thresholding for separately extracting the positive and negative frequency envelopes that are present in spectrogram. The proposed method does not make use of a deep learning algorithm.

In [12], a vision-based application is created that can offer sign language translation. The proposed method extracts temporal and spatial features from the video sequences. For spatial feature recognition CNN is used and a recurrent neural network (RNN) is applied to train on the temporal features.

In [13], a method is presented using deep convolution neural network (DCNN) to classify images of the letters and digits in ASL. The data set of 25 images from five different people were collected using a camera. An accuracy of 82.5% is achieved on the alphabet gestures, and 97% on digits.

Unlike the previous studies, which mainly focus on ASL letter or digit recognition, in this paper, we investigate recognition of ten essential hand gesture phrases including emergency signals. In an emergency situation, a first responder who may be unfamiliar with ASL can use the proposed system to quickly recognize emergency ASL phrases. The motion variations produced by hand gesture are captured by a microwave X-band Doppler radar transceiver. The captured signal is fed into a data acquisition device (DAQ) and using National Instrument (NI) LabVIEW SignalExpress software, the raw data is imported to a laptop for signal processing. We apply joint time-frequency analysis and observe the presence of micro-Doppler signatures in the spectrogram. From these observations, we notice that hand gestures contain unique micro-Doppler signatures. Based on unique micro-Doppler signatures, we build an ASL classification scheme that classifies important short phrases including emergency signals such as “Help me”, “Call 911”, “Danger”, “Don’t touch”, “Do you need help?”, “Call an Ambulance”, “How are you?”, “Nice to meet you”, “Yes” and “No”. To recognize the micro-Doppler signatures, we explore a DCNN algorithm, which is considered one of the most successful deep machine learning algorithms for image recognition [14], [15]. Spectrogram can be considered as an image in which case applying DCNN can serve well for the feature recognition purposes. From the captured data, we crop and collect the spectrograms of the different phrases. We then apply the DCNN algorithm on to the captured raw micro-Doppler spectrograms. With a fairly high accuracy DCNN algorithm classifies different ASL gestures based on spectrograms. Experiments are conducted using a total of 400 spectrogram images for 10 different gestures of which 80% is used for training purposes, and the remaining 20% is used for validation purposes. The experimental results reveal that the average validation accuracy is 87.5%. To further improve the recognition system a very deep learning algorithm VGG-16 [16] using transfer learning is explored, which raises the validation accuracy to 95%.

The rest of the paper is organized as follows. In Section II, we discuss micro-Doppler signatures generated by the sign language gestures, followed by the sign language gesture detection with Doppler radar in Section III. The deep convolution neural network algorithm is presented in Section IV. After illustrating experimental results in Section V, we draw the main conclusions in Section VI.

II. Modeling Micro-Doppler Signatures from Sign Language Gestures

Sign language gestures are produced by the motion of the speaker’s hands in a certain pattern. These motions can be captured when illuminated by a radar signal. When a radar device transmits a pure tone at carrier frequency f_c onto a person communicating using sign language, the reflected signal contains micro-Doppler effects centered around the f_c, due to micro-motion variations of hands.

The received Doppler signal as a function of time is modeled as [17]

$$s(t) = Ae^{j(2\pi f_c t + \beta \mathsf{sin}(2\pi f_\nu t))}, \tag{1}$$

where A is the reflectivity of the vibrating point scatterer, $\beta \mathsf{sin}(2\pi f_\nu t)$ is the time-varying phase change of the vibrating scatterer in which f_ν is the frequency of vibrating scatterer and $\beta = 4\pi D_\nu/\lambda$, D_ν is the amplitude of the vibration and λ is the wavelength of the transmitted signal.

Since (1) is a periodic function it can be expanded using Fourier series as

$$s(t) = A \sum_{n=-\infty}^{\infty} c_n e^{j2\pi(f_c + nf_\nu)t}, \tag{2}$$

where c_n is the Fourier series coefficient, which is expressed as

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j\beta \mathsf{sin}(2\pi f_\nu t)} e^{-jn2\pi f_\nu t} dt = J_n(\beta), \tag{3}$$

where $J_n(\beta)$ is the nth-order Bessel function of the first kind.

Substituting (3) into (2) yields

$$s(t) = A \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi(f_c + nf_\nu)t}. \tag{4}$$

Equation (4) represents a micro-Doppler frequency spectrum consisting of pairs of harmonic spectral components centered around the carrier frequency f_c. The spacing between the adjacent spectral lines is governed by f_ν. Since the phase of the reflected wave expressed in (1) is time-varying, the instantaneous frequency f_D, which represents the micro-Doppler frequency induced by the vibrations of the scatterer, can be expressed as

$$f_D = \frac{1}{2\pi}\frac{\varphi(t)}{dt} = \beta f_\nu \mathsf{cos}(2\pi f_\nu t) \tag{5}$$

$$= \frac{4\pi}{\lambda} D_\nu f_\nu \mathsf{cos}(2\pi f_\nu t). \tag{6}$$

The maximum micro-Doppler frequency change is $\frac{4\pi}{\lambda} D_\nu f_\nu$, which can be used to estimate the maximum displacement of a vibrating scatterer. The micro-Doppler caused by vibration is a sinusoidal function of time at the vibrating frequency f_ν. The hand gesture vibrations produce micro-Doppler perturbations centered around the carrier frequency f_c can be used for sign language gesture detection and classification.

III. Sign Language Gesture Detection with Doppler Radar

In this section, we analyze the possibility of detecting sign language gestures using Doppler radar. The hand gesture motion variations are captured as micro-Doppler signal using a microwave X-band Doppler radar HB100 transceiver, shown in Fig. 1. The captured signal is fed into the DAQ device with a sample rate of 1 ksps and the raw data is imported to a laptop for signal processing using NI LabVIEW SignalExpress software. The hand gesture extraction process flowchart is shown in Fig. 2.

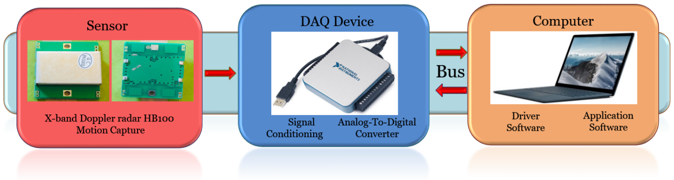


Fig. 1: Interfacing radar sensor with DAQ system.

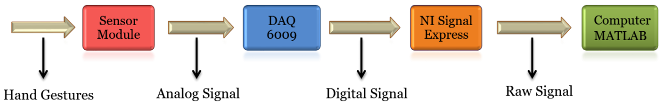
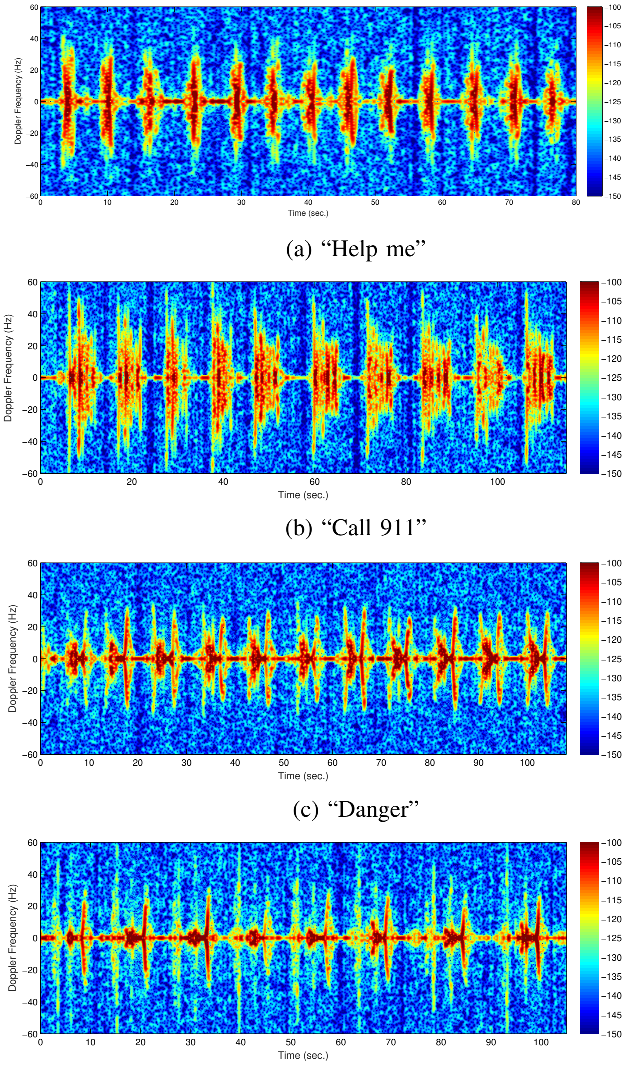


Fig. 2: Hand gesture micro-Doppler extraction process.

After removing the carrier frequency, we plot in Fig. 3 four different samples of ASL hand gesture spectrogram snapshots for a) "Help me", b) "Call 911", c) "Danger" and d) "Don't touch". As we can see from Fig. 3, each gesture has a unique spectrogram image characteristics and could be differentiated from each other even with a naked eye. The spectrogram, as an example, shown in Fig. 3a contains 13 samples of "Help me" ASL patters. The micro-Doppler variations are observed to be in the range from $0-50$ Hz. Each of these gestures in the spectrogram plot are cropped into images of size 100×100 pixels for signal processing in the DCNN algorithm.



(a) "Help me"

(b) "Call 911"

(c) "Danger"

(d) "Don't touch"

Fig. 3: Hand gesture spectrogram snapshots:

IV. Radar Micro-Doppler Classification with Deep Convolution Neural Network

A number of advanced CNNs algorithms have been developed for image classification [2], [3], [5], [7]–[9], [18], [19].

Figure 4 depicts a generalized network of CNN. The CNN algorithm extracts features from the training images and generates classifiers. The classifier weights are determined through the training process. The produced output $\boldsymbol{y}$, shown in Fig. 4, is compared with the input data $\boldsymbol{d}$ and the error information $\boldsymbol{e}$ is fed back to the algorithm to improve the classification process. In general, 80% of data is used for training purposes, and the remaining 20% for validating the CNN algorithm [21].

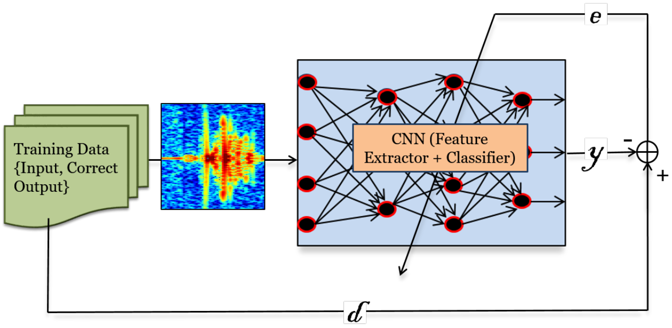

e
Training Data {Input, Correct Output}
CNN (Feature Extractor + Classifier)
y
-
+
d

Fig. 4: Generalized network of CNN.

To classify the different swimming styles, we implement a DCNN algorithm in Matlab as follows. The captured spectrogram images are first manually cropped to 100×100 pixel RGB images with ten classes, $1-10$. The input images undergo feature extraction network by first being processed by the convolution layer consisting of 8 convolution filters of size 20×20. The output from the convolution layer goes through the rectified linear unit (ReLU) function followed by the pooling layer, which employs max pooling process of 2×2 matrices. This process is repeated several times to create the output and train the machine with inherent features of the image. The output of the pooling layer is fed into a second convolution layer consisting of 16 convolution filters of size 10×10. Similarly, after passing the output through the ReLU function it undergoes the pooling layer with max pooling size of 2×2 matrices. Finally, it is passed through a third round of convolution layer consisting of 32 convolution filters of size 5×5 after which it is processed by the ReLU function and the pooling layer with max pooling size of 2×2 matrices.

The max pooling concept is demonstrated in Fig. 6. The stride is the sliding window operation, which is used in the convolution layer and in the max pooling operation in which case the stride is 2. Suppose $n\times n$ convolution is performed, the stride represents the movement by S elements with every step. If the stride is defined as 1 that means the convolution layer will move with sliding window of 1 pixel and move every third pixel by skipping the second pixel. Max pooling is a downsampling process where it selects the maximum value from each view. Since the spectrogram images contain sharp

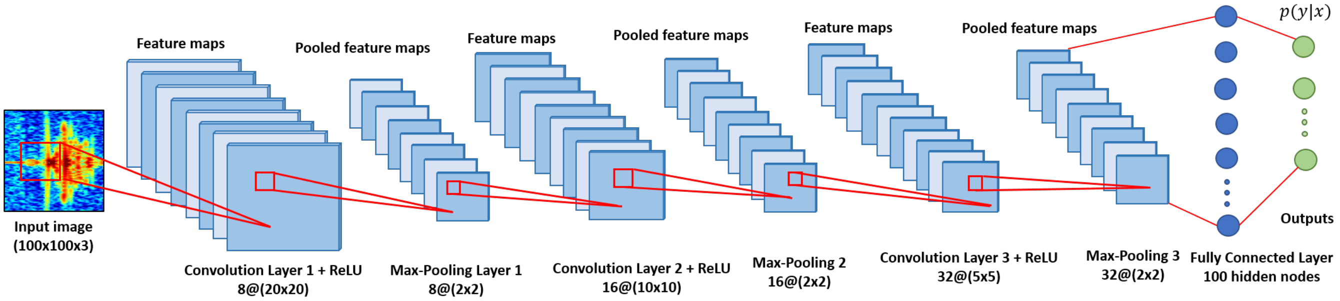
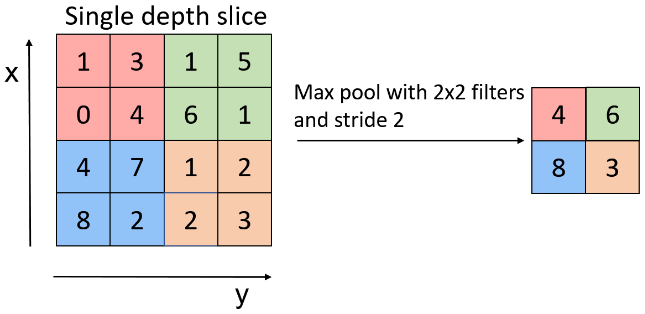


Fig. 5: Architecture of the DCNN algorithm implemented in Matlab.

edges max pooling instead of average pooling is used to extracts the most important features such as edges. The classifier network consists of a fully connected layer comprised of 100 hidden nodes, which produce a Softmax output that in turn is used for classifying the ten different ASL gestures. The architecture of the DCNN algorithm implemented in Matlab is shown in Fig. 5.



1	3	1	5
0	4	6	1
4	7	1	2
8	2	2	3

4	6
8	3

Fig. 6: Max pooling principle.

V. Experimental Results

In this section, we evaluate the performance of the proposed ASL classification scheme. The experimental setup of the testbed is shown in Fig. 7. Ten different types of ASL gestures are performed 40 times each and are captured by the microwave X-band HB100 radar transceiver. A total of 400 ASL gestures are gathered. Hand gestures are performed by an individual at a distance of approximately 80 cm from the radar sensor. The HB100 microwave sensor transmits pure tone at carrier frequency of 10.52 GHz. The reflected received sinusoidal signal is fed to the NI DAQ 6009 device that converts the analog signal to digital and feeds it to the NI LabVIEW SignalExpress software. The captured raw data is imported to Matlab to plot the spectrograms of the different ASL gestures. The 10 different gesture spectrograms are cropped and collected in a folder for classification purposes.

The proposed DCNN algorithm discussed in Section IV is implemented in Matlab to classify the ten ASL hand gestures. We used Dell Latitude E547 laptop with an 8th Generation Intel Core i7 processor for running the deep learning algorithm. Stochastic gradient descent (SGD) algorithm with momentum is applied to accelerate the learning rate. In the DCNN algorithm, the convolutions layers are defined with the batch normalization layers and the ReLU layers. The batch normalization layer helps to normalize the input layer by adjusting and scaling the activations, which can speed up the learning process. The ReLU layer captures interactions and non-linearities and can greatly accelerate the convergence of the SGD algorithm. The convolution layers of the DCNN algorithm can be changed according to the needs of experimentation.

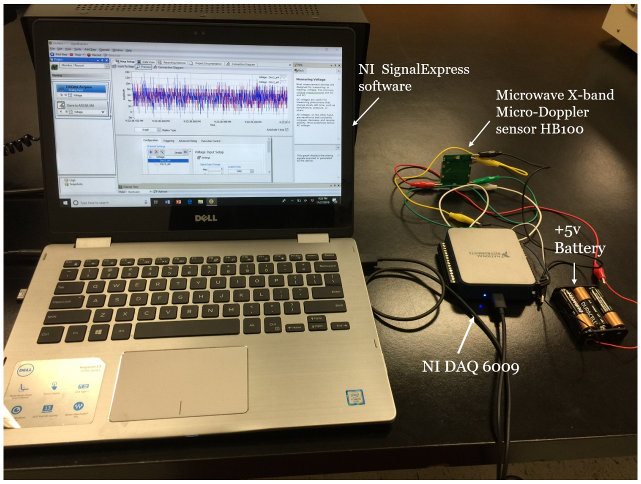
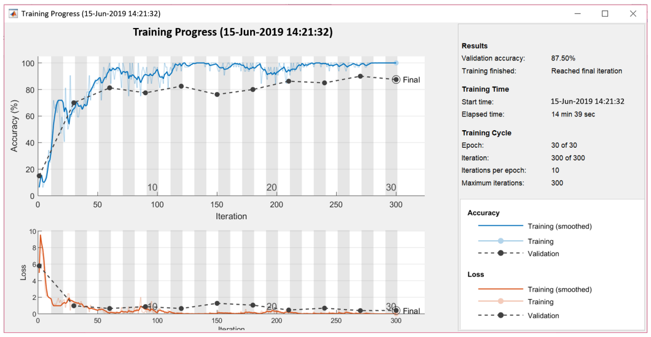


Fig. 7: Experiment setup.

The proposed DCNN algorithm is used to train those ASL gestures depicted by the spectrograms. A data set of 400 spectrogram images are used for the 10 different ASL gesture classification includes; “Help me”, “Call 911”, “Danger”, “Don’t touch”, “Call an Ambulance”,“How are you?”, “Nice to meet you”, “Yes” and “No”. Out of these 400 images 320 are used for training and the remaining 80 for validation purposes.

In Fig. 8, we plot the validation accuracy of the DCNN algorithm. The training accuracy graph is plotted in blue solid line and validation graph is plotted in black dotted line. The batch size used for the training purposes is selected to be 10, which is the number of iterations used for each epoch. A total of 30 epochs are used for data training, which resulted in 300 iterations. The experimental results reveal that the average



Training Progress (15-Jun-2019 14:21:32)

Training Progress (15-Jun-2019 14:21:32)

Results
Validation accuracy: 87.50%
Training finished: Reached final iteration

Training Time
Start time: 15-Jun-2019 14:21:32
Elapsed time: 14 min 39 sec

Training Cycle
Epoch: 30 of 30
Iteration: 300 of 300
Iterations per epoch: 10
Maximum iterations: 300

Fig. 8: Validation accuracy and packet loss graph of DCNN algorithm.

validation accuracy is 87.5%. The graph of the validation accuracy and training accuracy should be close to each other to overcome the overfitting problem. As we can see in Fig. 8, the validation loss decreases gradually with each iteration and the validation accuracy improves. However, it is important to note that since moderate sample data images were gathered in the experiment there is a gap about 12.5% between the training and validation accuracy.

To further improve the classification accuracy, we explore transfer learning. We apply VGG-16 algorithm, which is one of the well known trained DCNN algorithms. VGG-16 is developed by Karen Simonyan and Andrew Zisserman at the Visual Geometry Group (VGG) [23] and it is a very deep convolution network used for large-scale visual recognition. It is a pre-trained neural network, which has capability to train more than a million of images in 1000 categories. The VGG-16 algorithm consists of a total of 41 layers in which 16 are the convolution layers [16]. The architecture of VGG-16 is shown in Fig. 9. A preprocessing layer is included that takes the RGB image with pixels values in the range of $0 - 255$ and subtracts the mean image values.

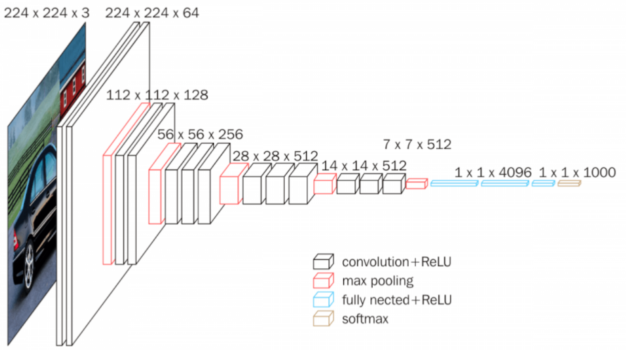


Fig. 9: VGG-16 architecture [22].

In Fig. 10, we show the fully connected 16 convolution layers that are used for training the data in the VGG-16 algorithm. The sizes of the convolution layers and max pooling layers used in the VGG-16 algorithm are 3×3 and 2×2, respectively. The stride used in the convolution layer is 1 and padding size is also 1. The stride size of max pooling layer is 2 and it does not use any padding. Padding is used by adding an extra bit that contains important information to protect the image from distortion. Since the image size is reduced in the pooling process it results in image distortion.

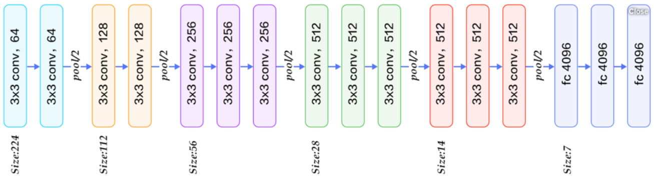
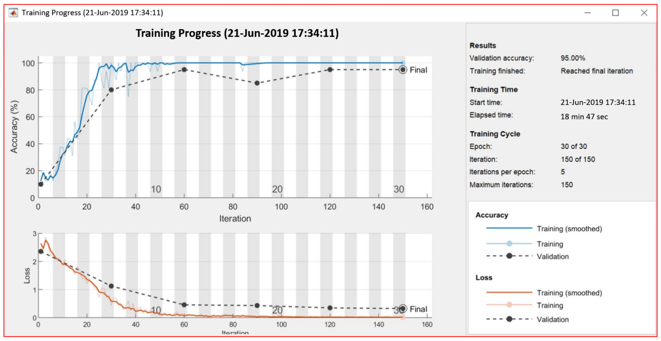


Fig. 10: Convolution layers of VGG-16 [23].

In Fig. 11, using the same experimental data, we train the VGG-16 algorithm using 30 epochs, 5 iteration per epoch, with a total of 150 iterations. The validation accuracy using the VGG-16 algorithm, as we can see in Fig. 11, raises to 95%. Comparisons of both results, DCNN and VGG-16 depicted in Figs. 8 and 11, respectively, VGG-16 provides higher validation accuracy and thus, it has better prediction capability than the DCNN algorithm. VGG-16 algorithm can take into account the small pattern variations and thus provides more accurate classification of the different ASL gestures experimented in this study.



Training Progress (21-Jun-2019 17:34:11)

Training Progress (21-Jun-2019 17:34:11)

Results
Validation accuracy: 95.00%
Training finished: Reached final iteration

Training Time
Start time: 21-Jun-2019 17:34:11
Elapsed time: 18 min 47 sec

Training Cycle
Epoch: 30 of 30
Iteration: 150 of 150
Iterations per epoch: 5
Maximum iterations: 150

Fig. 11: Validation accuracy and loss graph of VGG-16.

VI. Conclusion

In this paper, we investigated detection and classification of ten different American sign language (ASL) essential short phrase gestures including; “Help me”, “Call 911”, “Danger”, “Don’t touch”, “Do you need help?”, “Call an Ambulance”, “How are you?”, “Nice to meet you”, “Yes” and “No”. The gestures were captured by a X-band Doppler radar transceiver and extracted using National Instruments (NI) data acquisition (DAQ) device and LabVIEW SignalExpress software. The spectrogram images of hand gestures movements were trained and classified using deep convolution neural network (DCNN) algorithm and a very deep learning VGG-16 algorithm both of which were implemented in Matlab. We demonstrated that the DCNN algorithm can classify different ASL gestures based on spectrogram with a fairly high accuracy. The experimental results reveal that the average validation accuracy of DCNN and VGG-16 algorithms were 87.5% and 95%, respectively.

References

[1] Z. Chen, J.T. Kim, J. Liang, J. Zhang, and Y.B. Yuan, “Real-Time Hand Gesture Recognition Using Finger Segmentation,” *The Scientific World Journal*, vol. 2014, Article ID 267872, 9 pages, 2014.

[2] X. Li, Y. He and X. Jing, “A Survey of Deep Learning-Based Human Activity Recognition in Radar,” *Remote Sensing*, vol. 11, no. 9, 1068, May 2019.

[3] J. Wang, Y. Chen, S. Hao, X. Peng and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” *Pattern Recognition Letters*, vol. 119, pp. 3-11, Mar. 2019.

[4] R. Poppe, “A survey on vision-based human action recognition,” *Elsevier - Image and vision computing*, vol. 28, no. 6, pp. 976-990, Jun. 2010.

[5] Y. Sun, T. Fei, F. Schliep and N. Pohl, “Gesture Classification with Handcrafted Micro-Doppler Features using a FMCW Radar,” in *Proc. IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, Munich, Germany, Apr. 2018, pp. 1-4.

[6] H. Zhang, A.C. Berg, M. Maire, J. Malik, “SVM-KNN: Discriminative nearest neighbor classification for visual category recognition,” in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, New York, NY, USA, Jun. 2006, pp. 2126-2136.

[7] S. Zhang, G. Li, M. Ritchie, F. Fioranelli and H. Griffiths, “Dynamic hand gesture classification based on radar micro-Doppler signatures,” in *Proc. IEEE CIE International Conference on Radar (RADAR)*, Guangzhou, China, Oct. 2016, pp.1-4.

[8] M.A. Jalal, R. Chen, R.K. Moore and L. Mihaylova, “American sign language posture understanding with deep neural networks,” in *Proc. IEEE International Conference on Information Fusion (FUSION)*, Cambridge, UK, Jul. 2018, pp. 573-579.

[9] B. Garcia and S.A. Viesca, “Real-time American Sign Language Recognition with Convolutional Neural Networks,” *Convolutional Neural Networks for Visual Recognition*, Stanford University, (2), 2016.

[10] M.G. Amin, Z. Zeng and T. Shan, “Hand Gesture Recognition based on Radar Micro-Doppler Signature Envelopes,” in: *arXiv preprint arXiv:1811.12467*, Feb. 7, 2019, pp. 1-6. [Online]. available: https://arxiv.org/pdf/1811.12467.pdf

[11] S. Craw, “Manhattan Distance,” in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 639-639.

[12] K. Bantupalli and Y. Xie, “American Sign Language Recognition using Deep Learning and Computer Vision,” in *Proc. IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, Dec. 2018, pp. 4896-4899.

[13] V. Bheda, D. Radpour, “Using deep convolutional networks for gesture recognition in American sign language,” in: *arXiv preprint arXiv:1710.06836*, Oct. 18, 2017. [Online]. available: https://arxiv.org/ftp/arxiv/papers/1710/1710.06836.pdf

[14] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning”, in: *arXiv preprint arXiv:1603.07285v2*, Jan. 18, 2018. [Online]. available: https://arxiv.org/pdf/1603.07285.pdf

[15] I. Goodfellow, Y. Bengio and A. Courville, “Deep learning,” *MIT press*, Cambridge, MA, USA, Nov. 2016.

[16] K. Simonyan and A. Zisserman, “Deep Convolutional Networks for Large-scale Image Recognition”, in: *arXiv preprint arXiv:1409.1556*, Apr. 10, 2015. [Online]. available: https://arxiv.org/pdf/1409.1556.pdf

[17] V.C. Chen and H. Ling, “Time-frequency transforms for radar imaging and signal analysis,” *Artech house*, Norwood, MA, USA, 2002.

[18] H. Kulhandjian, N. Ramachandran, M. Kulhandjian, Claude D’Amours, “Human Activity Classification in Underwater using Sonar and Deep Learning,” in *Proc. ACM Intl. Conf. on Underwater Networks & Systems (WUWNet)*, Atlanta, GA, USA, Oct. 2019, pp. 1-5.

[19] Y. Shao, Y. Dai, L. Yuan, and W. Chen, “Deep learning methods for personnel recognition based on micro-Doppler features,” in *ACM Proc. of the 9th International Conference on Signal Processing Systems*, Auckland, New Zealand, Nov. 2017, pp. 94-98.

[20] V.C. Chen, F. Li, S.S. Ho and H. Wechsler, “Micro-Doppler effect in radar: phenomenon, model, and simulation study,” *IEEE Transactions on Aerospace and electronic systems*, 42(1), pp. 2-21, Jan. 2006.

[21] I. Rojas, G. Joya, A. Catala, “Advances in Computational Intelligence,” in *Springer - Proc. 13th International Work-Conference on Artificial Neural Networks, Part II, IWANN*, Palma de Mallorca, Spain, Jun. 2015.

[22] D. Frossard, “Model and pre-trained parameters for VGG16 in TensorFlow”, Jun. 2016, [Online]. available: https://www.cs.toronto.edu/~frossard/post/vgg16/

[23] K. Simonyan and A. Zisserman, “Visual Geometry Group (VGG)”, [Online]. available: https://www.robots.ox.ac.uk/~vgg/