Proceedings of the 2018 IISE Annual Conference
K. Barker, D. Berry, C. Rainwater, eds.

American Sign Language Alphabet Recognition Using Leap Motion Controller

Wenjin Tao, Ze-Hao Lai, Ming C. Leu
Department of Mechanical and Aerospace Engineering
Missouri University of Science and Technology
Rolla, MO 65409, USA

Zhaozheng Yin
Department of Computer Science
Missouri University of Science and Technology
Rolla, MO 65409, USA

Abstract

Recognition of American Sign Language (ASL) alphabet not only could bring benefits to the ASL users, but also could provide solutions for natural human-computer/robot interactions in many applications. In this paper, we propose a method for ASL alphabet recognition with use of a Leap Motion Controller (LMC). The skeleton data from the native LMC API is transformed by a skeleton module into a vector of the angle features. Meanwhile, two raw infrared-radiation (IR) images are captured and each of them is fed into a vision module using a Convolutional Neural Network (CNN) for visual feature extraction, which results in two feature vectors. Those three feature vectors are then fed into a fusion neural network to output the predicted label. An ASL alphabet dataset is established, on which the proposed model is evaluated. The results show that our proposed method achieves the prediction accuracies of 80.1% and 99.7% in the leave-one-out and the half-half experiments, respectively.

Keywords

American Sign Language, Leap Motion Controller, Convolutional Neural Network, deep learning, machine learning.

1. Introduction

American Sign Language (ASL) alphabet refers to 26 finger-spelled letters including 24 static ones and 2 dynamic ones '*J*' and '*Z*' [1]. Automatic recognition of these signs not only could bring benefits to the ASL users, but also could provide solutions for natural human-computer/robot interactions in a wide range of applications such as using gesture control devices. To tackle this challenging task, different kinds of sensors have been explored. Cyber gloves can directly sense the finger bending but are inconvenient for daily use [2, 3]. Microsoft Kinect, which was first released in 2010, has been applied in this task intensively because it is able to perceive the depth information [4, 5]. Leap Motion Controller (LMC) was first released in 2012 and it can provide the hand skeleton information. This information has been used for sign recognition tasks [6, 7]. Different machine learning methods have been applied, such as support vector machine, k-nearest neighbor, and random forest [4, 8]. Deep learning methods, such as Convolutional Neural Networks (CNN), have been prevalent recently for image recognition and Natural Language Processing (NLP) tasks [9].

In the present study, we use a LMC as the sensing device to capture human hands for sign recognition. An overview of the proposed model is illustrated in Figure 1. The LMC returns skeleton data and two infrared-radiation (IR) images, from which the angle features and the visual features are extracted by a skeleton module and a vision module, respectively. Subsequently, these extracted features are fed into a neural network classifier to make the final prediction. To evaluate the proposed model, a dataset is established on which the evaluation experiments are conducted.

The remainder of the paper is organized as follows. Section 2 presents the method of data acquisition and dataset establishment. Section 3 explains our proposed methodology for recognizing ASL alphabet. The experimental studies and results are discussed in Section 4. Finally, the conclusion is provided in Section 5.

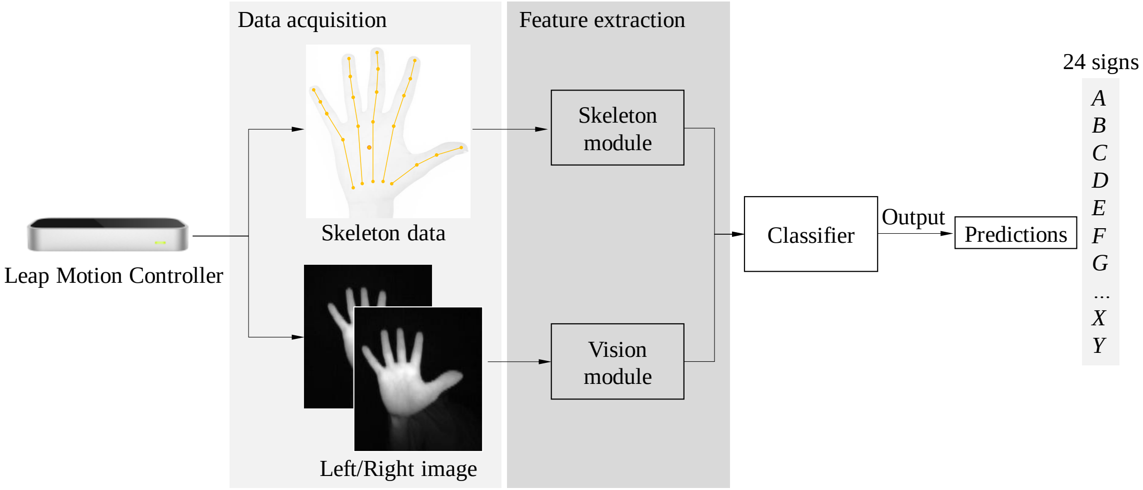


Figure 1: Overview of the proposed model

2. Data Acquisition

A Leap Motion Controller (LMC) developed by Leap Motion Inc. is chosen for data acquisition purpose due to its high precision, good portability, and low cost [10]. It is equipped with three infrared-radiation (IR) light-emitting diodes (LEDs) and two monochromatic IR cameras. These three LEDs emit pattern-less IR illuminations into an approximately hemispherical space, which will light up the hands appeared in this area. Then the illuminated hands can be perceived by the two cameras, which capture two IR images. The LMC provides a software to detect the hands in the images and recognize the hand skeletons such as bones, finger joints, and palm. It allows retrieving the skeleton data using an application programming interface (API).

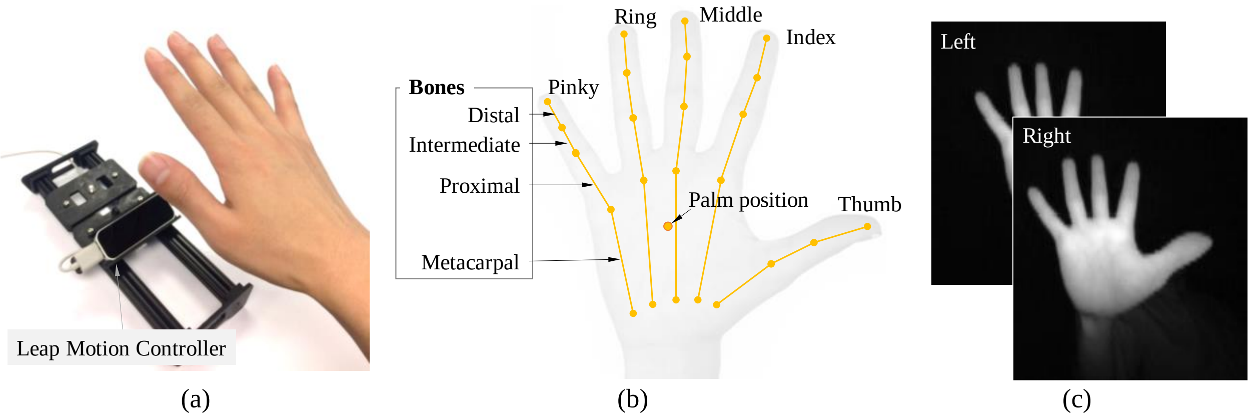


Figure 2: Illustrations of (a) the setup of data acquisition, (b) the skeleton data, and (c) the IR images from a LMC

To create our dataset of the 24 static ASL alphabet (the dynamic signs of '*J*' and '*Z*' are excluded), at present 5 subjects are recruited to perform these 24 signs. As illustrated in Figure 2(a), the subject is asked to perform a sign to the LMC using the right hand following the video instruction [11]. Meanwhile, the LMC returns two types of data, the skeleton data (see Figure 2(b)) and two IR images (see Figure 2(c)), which are recorded simultaneously at a rate of about 30Hz. To include more variations in the dataset, the subject is asked to keep adjusting the sign slightly in terms of hand distance to LMC, hand orientation, and finger posture. There are 450 frames recorded for each of the 24 signs during each recording. Then a calibration procedure is implemented on the captured IR images to address the issue of lens distortion. Since the hand region is illuminated by the IR LEDs, the undistorted images are then processed using a brightness threshold and cropped accordingly to remove the background.

3. Methodology

Figure 3(a) illustrates the overall architecture of our proposed model. It has a skeleton module and a vision module to extract features from the skeleton data and the two IR images, respectively. Then these features are fused together for classification.

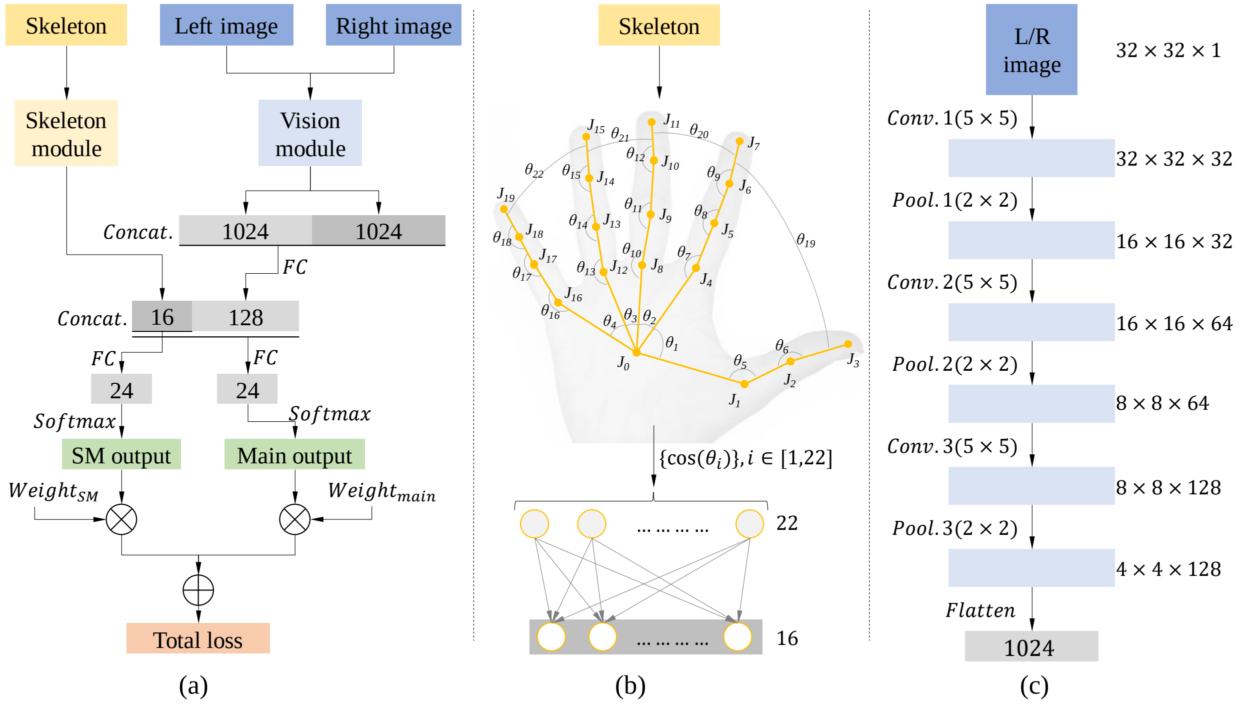


Figure 3: Structures of (a) the proposed model, (b) skeleton module, and (c) vision module. '*Concat.*', '*FC*', '*Conv.*', and '*Pool.*' denote the operations of concatenation, full connection, convolution, and pooling, respectively.

In the skeleton module (see Figure 3(b)), the positions of the 19 joints $\{J_1, J_2, \dots, J_{19}\}$ and the palm center J_0 are retrieved from the skeleton data. Then, their angle features are extracted as they are critical in defining different signs. There are 22 angles $\{\theta_1, \theta_2, \dots, \theta_{22}\}$, including angles between adjacent bones in the same finger and angles between adjacent fingers, selected and their cosine values are calculated as a 22-dimensional feature vector, which is fully connected to a layer with 16 neurons. Finally, the skeleton module outputs a feature vector with the dimension of 16.

The vision module is a Convolutional Neural Network (CNN) whose architecture is shown in Figure 3(c). Suppose there are N pairs of IR images X_i^{left} and X_i^{right}, $i \in [1, N]$. First, each of the images is scaled to the size of $32 \times 32 \times 1$ and its intensity is normalized to the interval $[0,1]$. Then three convolutional layers with the filter size of 5×5 and the increasing depth of 32, 64 and 128 are implemented to extract the visual features. A down-sampling layer is applied after each convolutional layer using the max pooling operation [12] with the filter size of 2×2 to reduce the spatial dimension (width and height). The third down-sampling layer $Pool.3$ returns a feature map with the size of $4 \times 4 \times 128$ which is flattened subsequently. Finally, the vision module outputs a feature vector with the dimension of 1024.

The fusion process is illustrated in Figure 3(a). First, the vision module generates a 1024-dimensional feature vector for each of the two IR images. Then these two vectors are concatenated into a 2048-dimensional vector on which a fully connected layer is applied forming a 128-dimensional vector. While the skeleton module generates a 16-dimensional vector, which is concatenated with the 128-dimensional vector forming a 144-dimensional vector. Finally, the fully connected layers with 24 neurons and the Softmax operations [12] are implemented on the 144 and 16-dimensional vectors, resulting in the main output and the SM output, respectively.

To train the model more effectively, two loss functions are considered, which are the cross entropies of the main and the SM outputs. The total loss is defined as the weighted summation of these two loss functions.

4. Experiments

We evaluate our method on the established ASL alphabet dataset, which has 24 signs performed by 5 subjects and 450 samples for each sign. Thus, there are 54,000 samples in total. Figure 4 shows examples of the 24 signs of each of the five subjects.

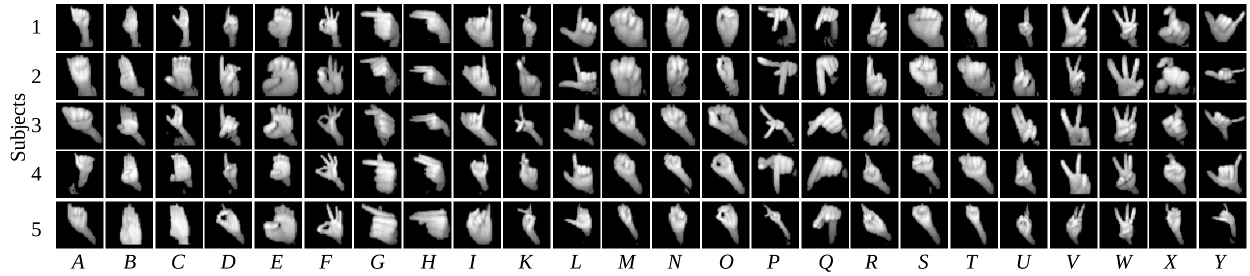


Figure 4: Examples of the 24 signs of each of the five subjects

The model described in Section 3 is created using TensorFlow [13] and Keras [14]. The leave-one-out and the half-half policies are conducted in the experiments for evaluation purpose. For the leave-one-out policy, the samples of one subject are left for evaluation, and the samples of the other subjects are used for training. For the half-half policy, one half of the samples is used for training and the other half is kept for evaluation. Recognition accuracy is used to quantify the classification performance, which is the percentage of correctly recognized signs.

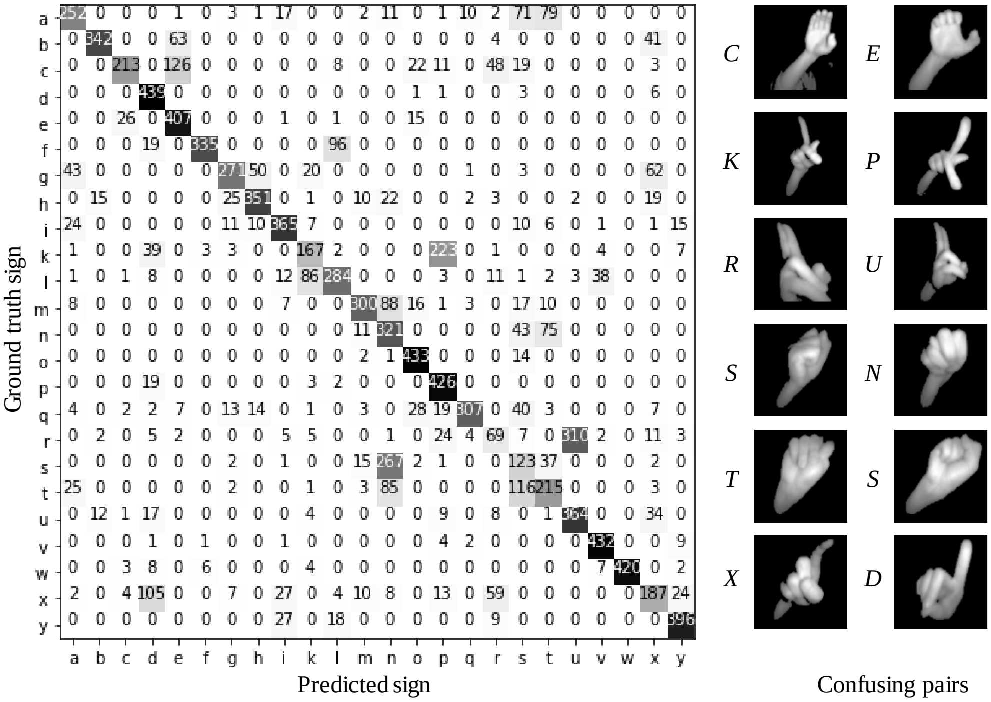
To evaluate the effectiveness of the proposed model, we compare the performance of four cases using different inputs: (1) single image (feed the vision module with only one image, discard the skeleton module); (2) two images (feed the vision module with two images, discard the skeleton module); (3) skeleton data (feed the skeleton module with the skeleton data, discard the vision module); and (4) the proposed model. The performance of the leave-one-out and the half-half evaluations in these four cases are summarized in Table 1. Case 2 has a higher accuracy than Case 1, which shows that using two images from two cameras helps to identify a sign. Case 3 has the lowest accuracy compared with the others, because the LMC cannot provide the correct skeleton data for some signs due to their high complexity and the finger occlusions. Overall, the proposed model has the highest prediction accuracies among the four cases in both the leave-one-out and the half-half evaluations, which are 80.1% and 99.7%, respectively.

Table 1: Prediction accuracies (%) of the leave-one-out (*loo*) and the half-half (*hh*) evaluations in different cases

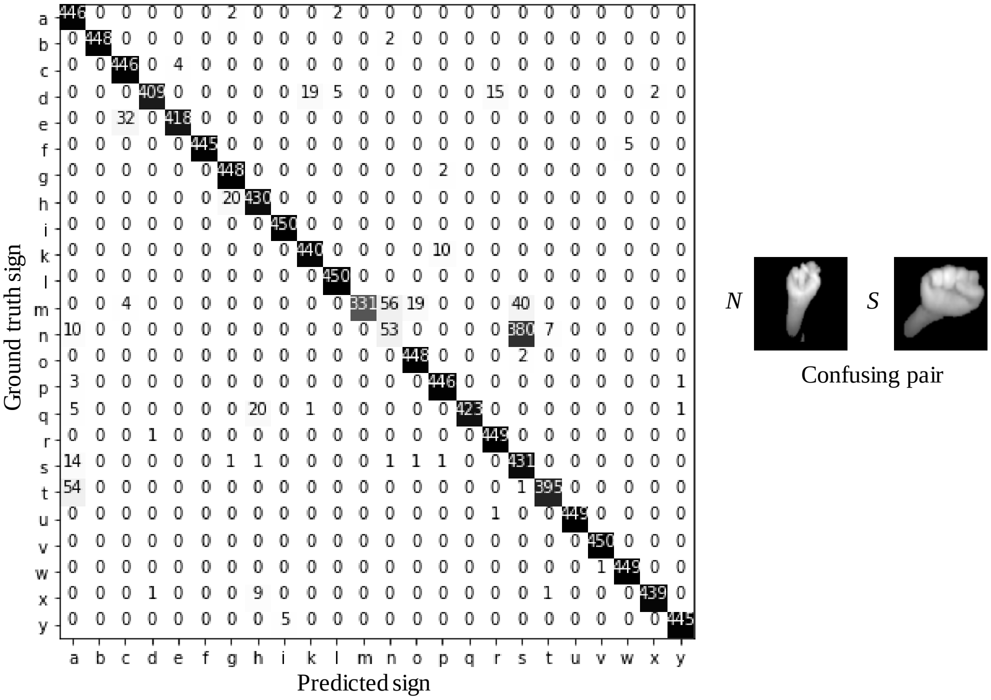
Cases	Single image		Two images		Skeleton data		Ours	
	loo	*hh*	*loo*	*hh*	*loo*	*hh*	*loo*	*hh*
1	87.9	-	86.1	-	35.2	-	87.2	-
2	75.4		78.5		22.8		81.1	
3	64.6		67.0		29.9		68.7	
4	90.6		93.0		39.7		91.2	
5	63.6		66.5		48.1		72.3	
Mean accuracy	76.4	99.4	78.2	99.6	35.1	66.7	**80.1**	**99.7**

As listed in Table 1, different subjects show different performance in the leave-one-out evaluations. The 4th subject has the highest accuracy of 91.2%, which is about 23% higher than the lowest one from the 3rd subject. As examples, the confusion matrices of the 3rd and the 4th subjects are shown in Figure 5. For the 3rd subject, there are some confusing pairs severely misclassified, such as '*C-E'*, '*K-P*', '*R-U*', 'S-N', '*T-S*', and '*X-D*', due to their high similarities (see Figure 5(a)). For the 4th subject (see Figure 5(b)), most of the signs are correctly classified except the most confusing pair '*N-S*' (i.e., there are 380 '*N*' misclassified as '*S*').

To solve these confusing cases, several approaches may be taken: (1) more subjects can be included to enhance the dataset; (2) data augmentation can be implemented to introduce more variations; (3) training of how to sign the ASL alphabet properly for the subjects is needed; (4) the skeleton and vision modules can be further explored to learn the most discriminative features; and (5) the fusion strategy can be improved for better performance.



(a)



(b)

Figure 5: Confusion matrices of the leave-one-out evaluations on the (a) 3rd and (b) 4th subjects

5. Conclusion

In this paper, we propose a method for American Sign Language (ASL) alphabet recognition using the skeleton data and two infrared-radiation (IR) images obtained from a Leap Motion Controller (LMC). A skeleton module is designed to extract angle features from the skeleton data, and a vision module, which is a Convolutional Neural Network (CNN), is developed to extract visual features from the two IR images. Then these two kinds of features are fused together with a neural network to output the final prediction. The proposed method compensates the weakness of the native LMC API in recognizing complex signs. An ASL alphabet dataset involving the 24 static signs performed by 5 subjects is created to evaluate our model. The experimental results on the established dataset demonstrate that our model achieves the prediction accuracies of 80.1% and 99.7% in the leave-one-out and the half-half evaluations, respectively.

Acknowledgements

This research work was supported by the National Science Foundation grant CMMI-1646162 on cyber-physical systems and also by the Intelligent Systems Center at Missouri University of Science and Technology. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. National Institute on Deafness and Other Communication Disorders, 2014. American Sign Language. URL: https://www.nidcd.nih.gov/health/american-sign-language/; [Online; accessed 15-January-2018].
2. Oz, C. and Leu, M.C., 2005, May. Recognition of finger spelling of American sign language with artificial neural network using position/orientation sensors and data glove. In International Symposium on Neural Networks (pp. 157-164). Springer, Berlin, Heidelberg.
3. Oz, C. and Leu, M.C., 2007. Linguistic properties based on American Sign Language isolated word recognition with artificial neural networks using a sensory glove and motion tracker. Neurocomputing, 70(16), pp.2891-2901.
4. Dong, C., Leu, M.C. and Yin, Z., 2015. American sign language alphabet recognition using microsoft kinect. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 44-52).
5. Nai, W., Liu, Y., Rempel, D. and Wang, Y., 2017. Fast hand posture classification using depth features extracted from random line segments. Pattern Recognition, 65, pp.1-10.
6. Lu, W., Tong, Z. and Chu, J., 2016. Dynamic hand gesture recognition with leap motion controller. IEEE Signal Processing Letters, 23(9), pp.1188-1192.
7. Kumar, P., Gauba, H., Roy, P.P. and Dogra, D.P., 2017. Coupled hmm-based multi-sensor data fusion for sign language recognition. Pattern Recognition Letters, 86, pp.1-8.
8. Chuan, C.H., Regina, E. and Guardino, C., 2014, December. American sign language recognition using leap motion sensor. In Machine Learning and Applications (ICMLA), 2014 13th International Conference on (pp. 541-544). IEEE.
9. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature, 521(7553), pp.436-444.
10. Leap Motion Inc., 2017. Leap Motion Controller. URL: http://www.leapmotion.com/; [Online; accessed 15-January-2018].
11. Youtube, 2013. The ASL Alphabet. URL: https://www.youtube.com/watch?v=tkMg8g8vVUo/; [Online; accessed 15-January-2018].
12. Goodfellow, I., Bengio, Y. and Courville, A., 2016. Deep Learning. MIT Press.
13. Google, 2015. TensorFlow. URL: https://www.tensorflow.org/; [Online; accessed 15-January-2018].
14. Chollet, F., 2015. Keras. URL: https://keras.io/; [Online; accessed 15-January-2018].