Classification with Sign Language Alphabet using CNN Convolutional Neural Networks

Mark Allen Cabutaje
BS in Computer Science
Technological University of the Philippines
Ayala Boulevard, Ermita, Manila
Philippines
markallen.cabutaje@tup.edu.ph

Kenneth Ang Brondial
BS in Computer Science
Technological University of the Philippines
Ayala Boulevard, Ermita, Manila
Philippines
kenneth.brondial@tup.edu.ph

Alyssa Franchesca Obillo
BS in Computer Science
Technological University of the Philippines
Ayala Boulevard, Ermita, Manila
Philippines
alyobllo05@gmail.com

***Abstract*— Sign language is the most effective way for deaf people to communicate. Numerous sign languages are built of various gestures generated by varied hand shapes, hand positions, or color. This project proposes a Hand Gesture Image Recognition for Sign Language using Convolutional Neural Network, that can recognize and predict a certain letter, using an image training model. The dataset gathered from Kaggle, was divided into two sets: training and test set, we will use them to test and train to our system, so that the prediction rate will be more accurate and reliable. Through learning the basic sign language alphabet, it will help even an ordinary person somehow communicate to people with hearing loss. This project aims to act as a bridge for deaf individuals to converse more effectively with non-sign language speaker.**

Keywords—Deep learning, convolutional neural network, sign language, Human Computer Interaction, Hand gesture, Image processing, Pattern classification, Image sequence, Binary images, Tensor Flow

I. INTRODUCTION

Sign language is commonly used by those who are unable to speak and hear and those who can hear but cannot communicate. Numerous sign languages are built of various gestures generated by varied hand shapes, hand motions, body orientations, or facial expressions. The deaf community uses these gestures to communicate their thoughts. However, these gestures are always restricted inside the deaf community; normal individuals would never attempt to learn sign language. This creates a significant communication barrier between deaf individuals and normal individuals. Sign language interpreters are in high demand because of the dramatic rise in the number of people who use sign language for communication. Unfortunately, this kind of translator is hard to come by and prohibitively costly. Automated sign language recognition systems were developed as a consequence of this, allowing people to communicate using sign language without the assistance of interpreters. Human-computer interaction in such systems may aid the deaf community's growth.

There are numerous types of sign language recognition, such as the Hardware-based systems and vision-based systems. These are the two main types of sign language recognition systems. Hardware-based systems needs the user to wear specific equipment in order to extract characteristics defining the hand sign. For instance, a cyber glove is a device that can be used to extracts hand characteristics such as direction, movement, and color. It is commonly utilized in the identification of sign languages. The other common type of sign language recognition, is the vision-based systems. It primarily uses a digital image processing methods to extract characteristics and detect signs. The technique suggested in this project is a vision-based approach that does not need the user to wear any specific devices.

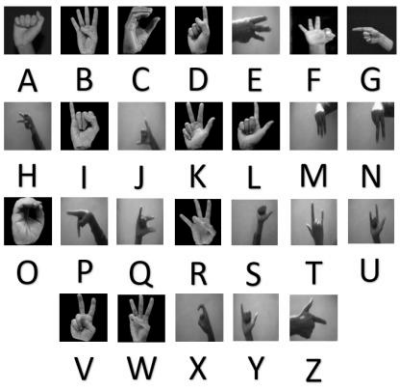


Figure 1. Sign Language Alphabet

Sign language image recognition is an important research area since there are a lot of challenges in developing an automatic recognition system. In implying this kind of mindset, we use the Convolutional Neural Network (CNN), in order for us to train and classify picture/s from the dataset that is available on the Kaggle. In this study, the user must take photo/s of the hand gesture using a web camera, and the system must anticipate and display the result, together with its level of confidence, from the processed image. To identify the hand gesture, the photos are processed in a number of phases that comprise several computer vision methods, specifically the grayscale conversion. We have a good level of accuracy in recognizing all sign gesture alphabets. Our model has a remarkable accuracy of more than 84%.

The arrangement of this paper are as follows; Section II provides the domain information for the algorithm used, which is the Convolutional Neural Network; Section III discussed the methodology in developing, validating, and evaluating using evaluation metrics; Section IV presents the results of the trained images from the algorithm used; and Section V shows the conclusion based on the results, and the relevant studies.

II. BACKGROUND

A. Convolutional Neural Network (CNN)

Deep Learning algorithms are built to simulate the operation of the human brain. These algorithms are models of deep neural networks, which include many hidden layers. [1] In particular, the Deep Convolutional Neural Network has demonstrated superior performance in image representation and classification, compared to conventional machine learning approaches. [2] Convolutional Neural Networks is one of the most effective deep learning algorithms. Using CNN, it can extract possible features for the classification model from a wide variety of images. Additionally, it generates an activations technique that serves as the input picture for each layer. Convolution occurs layer by layer. However, only a few layers within a CNN are sufficient for picture feature extraction. Convolutional neural networks are deep learning algorithms that can train enormous datasets with millions of parameters as input in the form of pictures and convolve them with filters to get the required outputs. CNN models are created elaborated used in this project, to test its performance using picture recognition and detection of datasets. [3] CNN have multiple layers; including convolutional layer, non-linearity layer, pooling layer and fully-connected layer. The CNN has an excellent performance in machine learning problems. Specially the applications that deal with image data, such as largest image classification data set (Image Net), computer vision, and in natural language processing (NLP) and the results achieved were very amazing.

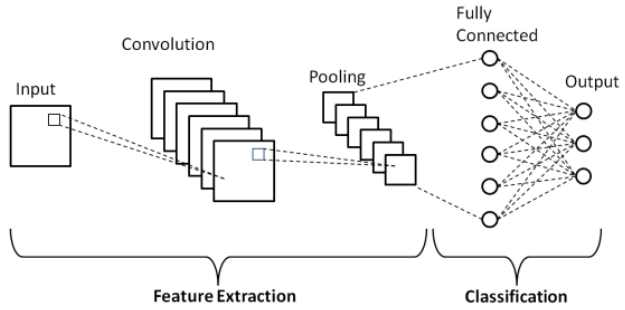


Figure 2. Convolutional Network Architecture

To further discuss the CNN, here is a much clearer way on how to understand this algorithm. CNN applies a series of filters to the raw pixel data of an image to extract and learn higher-level features, which the model can then use for classification. [4] CNNs contains three components:

Convolutional layers, which apply a specified number of convolution filters to the image. For each subregion, the layer performs a set of mathematical operations to produce a single value in the output feature map. Convolutional layers then typically apply a ReLU activation function to the output to introduce nonlinearities into the model.

Pooling layers, which down sample the image data extracted by the convolutional layers to reduce the dimensionality of the feature map in order to decrease processing time. A commonly used pooling algorithm is max pooling, which extracts subregions of the feature map (e.g., 2x2-pixel tiles), keeps their maximum value, and discards all other values.

Dense (fully connected) layers, which perform classification on the features extracted by the convolutional layers and down sampled by the pooling layers. In a dense layer, every node in the layer is connected to every node in the preceding layer.

B. CNN Advantages

Using CNN as an algorithm, it can; learn accurate pattern and insights from the provided data (but, it depends on how well structured, clean or feature engineered the data is), one can tune the network to achieve better and accurate results, and can provide better outcomes than other machine learning algorithms if tuned better and feeded a good amount of data.

C. Technologies Used

Python [5]

Python is an interpreted high-level general-purpose programming language. Its language elements and object-oriented approach are intended, to assist programmers in writing clear, logical code for small and large-scale projects.

OpenCV [6]

OpenCV is a short form for Open-Source Computer Vision Library, features programming languages, like C++, Python, and Java interfaces and works with a variety of platforms including windows, Linux, Mac OS, iOS, and Android. OpenCV is intended to enhance computational efficiency while putting a significant emphasis on real-time applications.

Tensor Flow [7]

Tensor Flow is a free and open-source software framework that allows high-performance numerical computing. It offers significant support for machine learning and deep learning, and its versatile numerical computation core is employed in a variety of other fields of science.

Media Pipe [8]

MediaPipe Hands is a solution for high-fidelity hand and finger tracking. Machine learning (ML) is used to deduce 21 3D landmarks of a hand from a single shot.

III. EXPERIMENTS

Hand gesture recognition is a way of solving numerous challenges and making life easier for humans. Machines' ability to comprehend human actions and their meaning may be used to a wide range of situations. It is one area of research that has peaked our curiosity. The methodologies

are broken down into stages: data collecting, pre-processing, the evaluation metrics, architecture, and the training.

A. Data Collecting

Data Set and Parameters

The system was created using a Python framework, that took one image of a whole user with the sign, before transmitting it to the server. It is created to acquire a decent dataset, which collected one photograph of random size picture every 1 second. Regardless with the distinct backdrop, lighting, and sign viewpoint, we have successfully used all the data we have gathered. A total of 39,000 photos were gathered for 26 classes, from the alphabet. Table 1 displays the number of classes and the number of photos used.

No. of Classes	Classes	No. of Image per class	Image size and file type	Color mode	Division
26	**A, B, C, D, E, F G, H, I J, K, L M, N, O P, Q, R S, T, U V, W, X Y, Z**	**Consists of 1500 images per letter**	**Diff image size, .jpg**	**Gray scale**	**Training : 64%, Validation: 16%, Test: 20%**

Table 1: Classification of the Dataset

Developing a Data Model for our Image Data

In order to resolve an image classification problem, we require our data in a certain format. 64% from the dataset will be for the train set, the 16% will be for the validation, and last the remaining 20% will be for the test.

Our model will be trained using the photos in the training set, and the label predictions will be made using the images in the testing set. If the data does not already exist in the format mentioned above, we will have to convert it (otherwise the predictions will be fairly useless).

B. Pre-processing

We initially trained and tested some self-generated dataset of images we took ourselves. Since our dataset was not constructed in a controlled setting, it was especially prone to differences in light, skin color, and other differences in the environment that the images were captured in.

Data augmentation is applied, in which it shifts images both horizontally to an extent of 0.1 of the original dimensions randomly, in order to numerically increase the size of the dataset and to add the robustness needed for a deep learning approach, and in order also to avoid overfitting.

Augmentations	
Rescale	0.1 / 255
Horizontal Flip	True
Rotational Range	0.1

Table 2: Dataset Augmentations

C. The evaluation metrics

Tensorflow

As it is mentioned in this paper, we used tensor flow libraries and with Keras as our backend. The method we user both allows high-performance numerical computing. It offers significant support for machine learning and deep learning, and its versatile numerical computation core is employed in a variety of other fields of science, to perform, training and validate the network/s.

Layer	Units	Kernel Size	Activation	Misc
Convolutional	16	3x3	ReLU	With padding
Max Pool 2D	-	2x2	-	-
Convolutional	32	3x3	ReLU	With padding
Max Pool 2D	-	2x2	-	-
Convolutional	64	3x3	ReLU	With padding
Max Pool 2D	-	2x2	-	-
Fully Connected	128	-	ReLU	-
Dropout	-	-	-	20%
Fully Connected	26	-	Softmax	-

Table 3: Network 1

D. Training

The photos in the dataset will be input into the Convolutional Neural Network in 15 epochs using one of the most common optimizer, aside from SGD, which is the Adam optimizer. To avoid overfitting, the training will use early stopping, in which the model fitting will immediately stop the training when it no longer learns. Holdout cross validation will also be used to validate the model. The accuracy of the model will be used as a criterion to compare the performance of various model setups.

E. Architecture

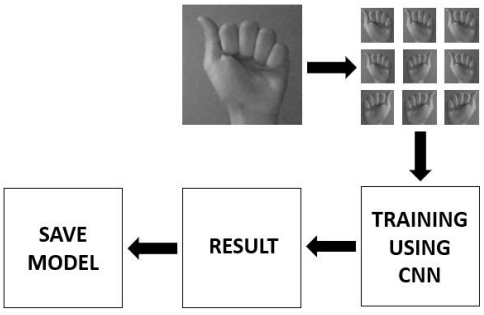


Figure 3: Network Architecture [Training Phase]

The architecture of our program, when in training phase is as follows: first we input the raw data in grayscale picture from the dataset, after that, it will go to the process of data augmentation. After the image did successfully go through data augmentation it will undergo to the CNN training method, wherein the program will produce a result. After getting the result, it will be saved to our program and this will benefit for our program to learn more about sign language recognition, when it is implemented.

The architecture of our system, when it is implemented after the dataset have been trained is as follows; first, from the inputted raw data, it will be converted into grayscale, after that, it will go to the process of predicting by the system. And this is where the result comes in, and it will predict what letter in the alphabet, specifically on the dataset that we trained about sign language, together with the level of its confidence.

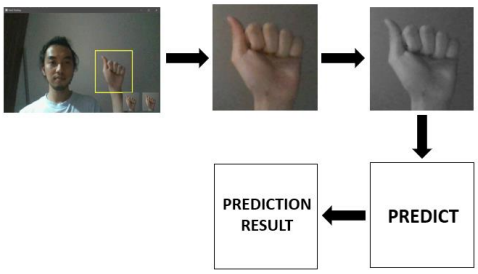
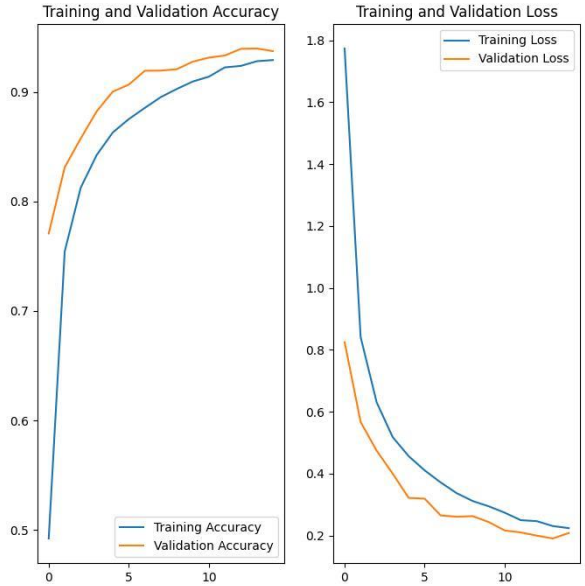


Figure 4: Network Architecture [Implementing Phase]

IV. RESULTS

A. Training Validation and Accuracy

The model peaked at 15th epoch and reached training accuracy of 92% and 93% on validation. The table below shows the training accuracy and validation accuracy of the model configurations we have produced.



Training Accuracy	Validation Accuracy
92%	93%

Table 4: Rate of Accuracy

B. Confusion Matrix

[9] With the use of the confusion matrix, the following formula is used to calculate accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positive (TP): The predicted value matches the actual value. The actual value was positive and the model predicted a positive value

False Positive (FP): The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value

True Negative (TN): The predicted value matches the actual value. The actual value was negative and the model predicted a negative value

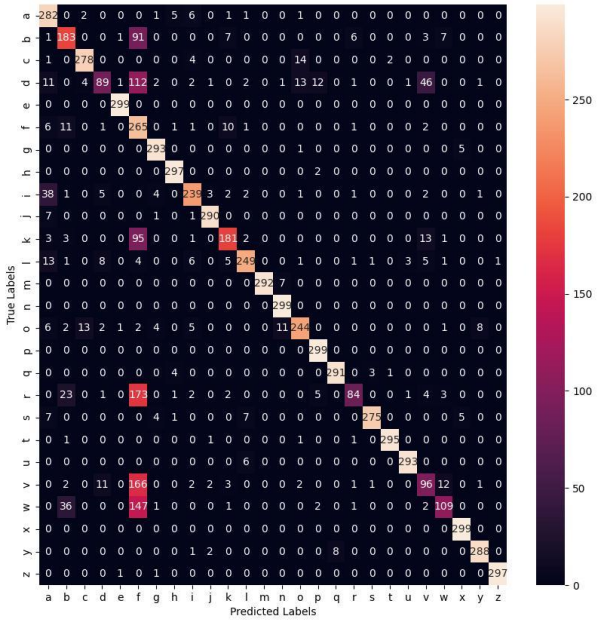
False Negative (FN): The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value

For multiclass classification, accuracy needs to be averaged.

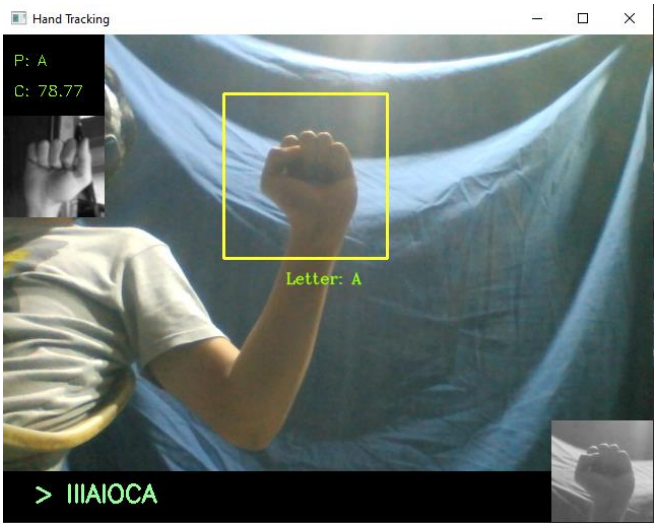
The average accuracy would be:

$$\sum_{i=1}^{k} \frac{TP + TN}{TP + TN + FP + FN}$$

where k = the number of classes.



C. *Trained System*



After the datasets have been trained to the system, it can be seen through the upper-left part. At the top of it, we can see the letter 'P', which shows the predicted letter by the system. Under it, is the level of its confidence, whether the system predicts that the letter that it predicted may be true or not. The system can detect the hand (in RGB mode) of a certain user within a distinct background. It will only get the desired image within the box it contains. The user can see on the lower right, the captured sign gesture from the user. On the lower part of the system, the user can generate a word base on the predicted gesture by the system, using the backspace as enter. When exiting, the user can type 'q' and it the system will exit.

V. RELATED WORK

Using Deep Convolutional Networks for Gesture Recognition in American Sign Language (Bheda et al, 2017), utilized a deep convolutional network to categorize ASL using alphabets and digits for specific CNN models. There were three cascaded convolutional layers and one max-pooling layer in the proposed network. Before connecting to the output layer, there are two hidden layers with dropout. An accuracy of 82.5 percent was reported. [10]

A real-time ASL recognition system used a Max-pooling convolutional neural networks for vision-based hand gesture recognition. It is primarily for the real colouring images from a PC camera. The model consists of ASL recognition model to categorize a total of 26 letters, including (J & Z). (Nagi, 2011) It was built to contain a wide diversity of attributes like different lightings, skin tones, backgrounds, and a wide variety of situations. The experimental results achieved a high accuracy of about 98.53% for the training and 98.84% for the validation. As well, the system displayed a high accuracy for all he datasets when new test data, which had not been used in the training, were introduced. [11]

Image processing with neural network (Michael et al, 2002), methods for recognizing bodies and body parts have been widely studied and a diverse set of applications created. Neural Networks (NN) are often used as recognition models for image processing. In this paper, he presented a basic NN model, and some variations applied to different areas. It primarily focuses on a specific type of network, Convolutional Neural Network (CNN), a popular NN in image processing because of its success recognizing local features. The recognized features are often used as input to NUI systems when interaction is via a user's body/body parts. This way of interacting avoids the use of physical control devices, allowing natural communication with the computer. The focus of NUI is to develop methods to provide effective user experience when interacting with the body directly and reduce ambiguity. [12]

American sign language numerals recognition from depth maps using artificial neural networks (Beena M.V. et al, 2017), developed a system to recognize American Sign Language (ASL) from the depth images of Kinect sensor. The system has trained using 1000 images of each numeral signs. The algorithm extract features from the block processed images and trained using the Artificial Neural Network (ANN) and obtained an accuracy of 99.46% for the depth images. The system has been trained on GPU for the faster execution. As an extension to the work uses Convolutional Neural Network [2017-2] (CNN) with SoftMax classification for 33 static symbols of Kinect depth images. The implementation shows that while the number of classes increases the handcrafted feature are become insufficient for classification purpose. The CNN structure is capable of learning from the given training set, and it will outperform the accuracy related to other traditional methods [13]

Human-computer interaction (Dix, 2016), Human-computer-interaction referred to as HCI is an interacting interface between humans (users) and machines (computers). Through HCI, humans and computers interact with each other in a novel way. Nowadays, it's a fascinating research field, which is focused on the designs and uses of computer technology and most particularly, the interacting interfaces between humans and machines. As day-by-day HCI technology has been remarkably expanded, research scope has been also raised up with the changes of technology. [14]

ImageNet classification with deep convolutional neural networks (Krizhevsky et al, 2012), Convolutional Neural Network (CNN), consists of one or more fully connected convolutional layers as standard multilayer neural network. CNN architecture is designed for handling 2D images efficiently. In this proposed model, CNN is used for extracting features in an automated way as its outcome is very satisfiable. Alternatively, CNN has several dynamic parameters to train up the machine easily. [15]

IV. CONCLUSION

Based on the findings reported in the preceding section, we can conclude that our Deep Learning-based system, correctly identifies diverse hand gesture photographs with sufficient confidence by using the test dataset, we were able to achieve an 82% accuracy as our final result.

There are a sufficient number of photos, which helped strengthened our model. But, the challenge, that we most likely faced, is the different images that we have been acquire have quality of the photo has a lot of dimensions/orientation, and some photos blends with its backdrop, in which our model quite had a difficulty. Furthermore, because to its abstractions, a deep learning model is extremely difficult to explain.

This project gave insight into the problem's constraints. We determined that the dataset used to train and assess the system must include enough gesture variants to allow for the generalization of each letter. We improved the average approach to assessing whether a dataset has sufficient variance in gestures and constructed a Convolutional Neural Network (CNN) for hand gesture recognition in pictures of Sign Language letters.

Additional classes will be added to the collection of gestures, and the optimal configuration will be determined using evaluation methods. Our objective is to derive a relationship between the number of classes and the optimal configurations from the relationship between the number of classes and the design of the CNN so that the system can develop a successful model for each case.

V. REFERENCES

[1] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 278-282, doi: 10.1109/ICSCCC.2018.8703316.

[2] A. Alani, G. Cosma, A. Taherkhani and T. M. McGinnity, "Hand gesture recognition using an adapted convolutional neural network with data augmentation," 2018 4th International Conference on Information Management (ICIM), 2018, pp. 5-12, doi: 10.1109/INFOMAN.2018.8392660.

[3] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[4] T. Mantecón, C.R. del Blanco, F. Jaureguizar, N. García, "Hand Gesture Recognition using Infrared Imagery Provided by Leap Motion Controller", Int. Conf. on Advanced Concepts for Intelligent Vision Systems, ACIVS 2016, Lecce, Italy, pp. 47–57, 24–27 Oct. 2016. (doi: 10.1007/978–3–319–48680–2_5)

[5] "Python." [Online]. Available: https://en.wikipedia.org/wiki/Python_(programming_language)

[6] "OpenCV usage documentationhttp://docs.opencv.org."

[7] "TensorFlow" [Online]. Available: https://en.wikipedia.org/wiki/TensorFlow

[8] Bhandari, A. (2020). "Confusion Matrix". Retreived from: analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/#:~:text=A%20Confusion%20matrix%20is%20an.by%20the%20machine%20learning%20model.&text=The%20rows%20represent%20the%20predicted%20values%20of%20the%20target%20variable

[9] Bheda, V., Radpour, D. (2017) "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language". Retrieved from https://arxiv.org/ftp/arxiv/papers/1710/1710.06836.pdf

[10] J. Nagi, F. Ducatelle, A. G. Di Caro, D. Cirean, U. Meier, A. Giusti, L. M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. 2011 IEEE International in Signal and Image Processing Applications (ICSIPA), Conference on, pp. 342-347 (2011)

[11] Petersen M., Ridder, D., Handels, H. (2002). Image processing with neural networksa review. Pattern recognition, 35(10):2279–2301, 2002.

[12] M.V. Beena and M.N. Agnisarman Namboodiri, ASL Numerals Recognition from Depth Maps Using Artificial Neural Networks, Middle-East Journal of Scientific Research 25 (7): 1407-1413, 2017,ISSN 1990-9233.

[13] A. Dix, "Human-computer interaction," in Encyclopedia of Database Systems, Springer US, pp. 1327–1331, 2016.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in proceedings of the 25th International Conference on Neural Information Processing Systems, pp. 1097–1105, Dec. 2012.