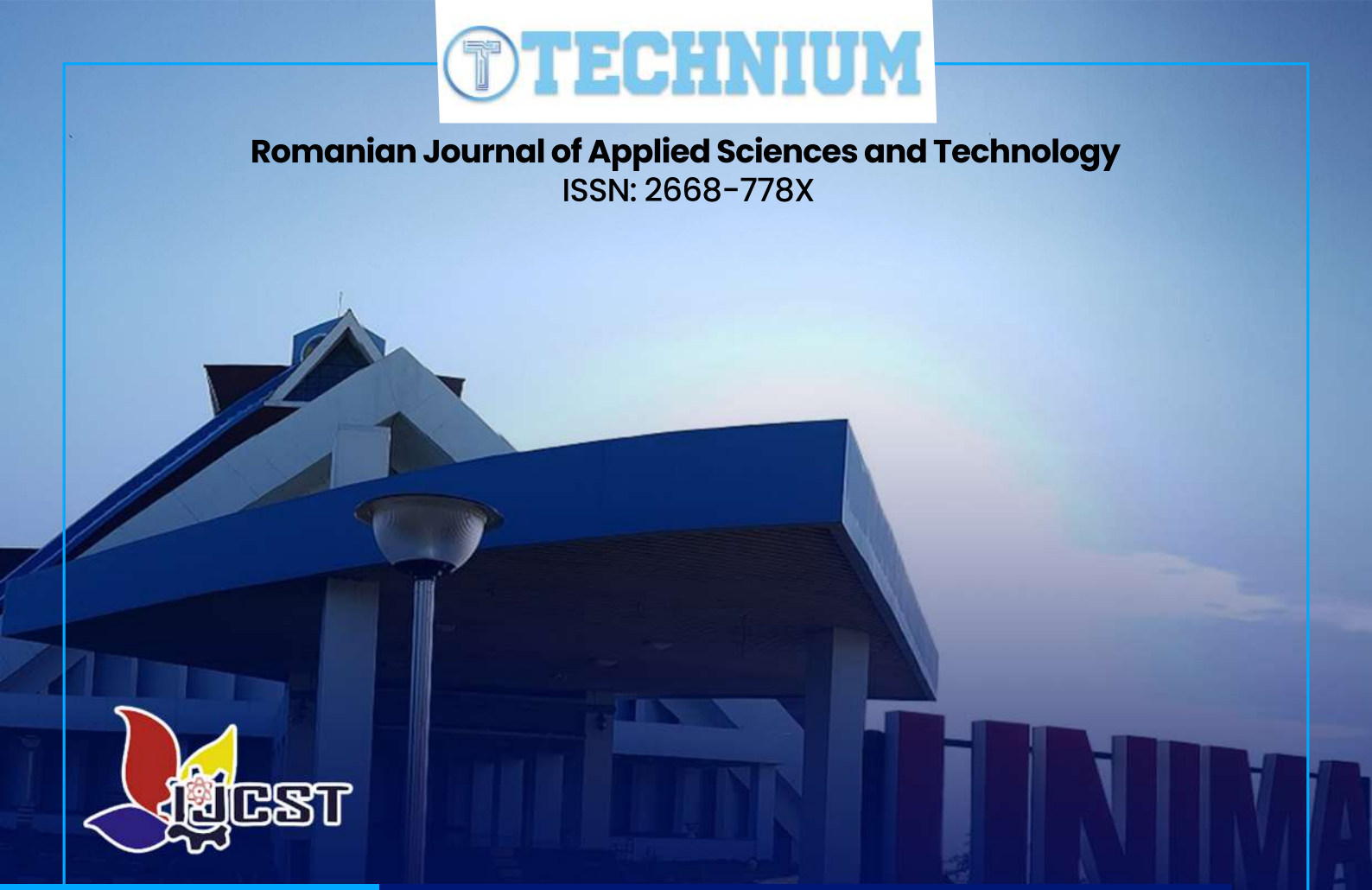
TECHNIUM

Romanian Journal of Applied Sciences and Technology
ISSN: 2668-778X



vol. 16 / 2023

The 7th International Conference on Science Technology

organized by
Faculty of Social Science and
Law Universitas Negeri Manado and
Consortium of International Conference
on Science and Technology

The Innovation Breakthrough in Digital and Disruptive Era

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 143-149, October, 2023
ISSN: 2668-778X
www.techniumscience.com

Object Localization and Detecting Alphabet in Sign Language BISINDO Using Convolution Neural Network

Yisti Vita Via[1*], *Wahyu* S. J. Saputra[2], *Mohammad* Idham Fachrurrozi[1], *Eva* Yulia Puspaningrum[1], *Fetty* Tri Anggraeny[1], and *Salamun* Rohman Nudin[3]

[1]Informatics, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia
[2]Data Science, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia
[3]Informatics Management, Vocational Program, Universitas Negeri Surabaya, Surabaya, Indonesia

Abstract. The BISINDO sign language is used to help deaf and mute people communicate with other people. However, not everyone is able to understand the meaning of this sign language. A system that implements artificial intelligence methods is created to solve this problem. The system uses a Convolution Neural Network algorithm with object localization techniques to detect and classify the alphabet in each form of the BISINDO finger signal. The Region Convolution Neural Network (RCNN) algorithm is used to process object localization and the CNN algorithm will perform classification process. This system is trained using 64 training data and tested using 16 test data for each type of alphabet. The results of the system testing that have been carried out are able to provide excellent accuracy values, which are above 90 percent for a training epoch of at least 50. These results produce an accuracy of 90.10% and 97.33% respectively.

* Corresponding author: yistivia.if@upnjatim.ac.id

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 143-149, October, 2023
ISSN: 2668-778X
www.techniumscience.com

1 Introduction

Social interaction is a human activity in dealing with other humans. These activities can occur when humans make social contact and communicate [1]. In the continuity of social interaction the use of language is the most meaningful symbol (significant symbol), through these symbols humans can not only interact with each other but can also interact with themselves in the process of thinking [2]. Communication is the key in socializing, so there is a social process in which each individual uses certain symbols in order to create and interpret meaning in the surrounding environment [3].

Communication is a means of conveying information with the process of giving and receiving various meanings between two or more people [4]. This is in line with what was expressed by Joseph A. Devito in his book entitled "The interpersonal communication book" which states that interpersonal communication is a process of sending messages by one person and receiving messages by other people or a small group of people, with various impacts and opportunities to provide immediate feedback [5]. So that the type of communication that has a high frequency is interpersonal communication. Therefore, it is not surprising that many people think that interpersonal communication is easy to do [6].

In its development in society, someone who can hear is usually called a "hearing community", the main means of direct communication using spoken language. However, some other community groups use spoken language may not be effective when communicating. Another community group is one of the deaf people with a slightly special condition where they do not have the ability to hear because they do not have a sense of hearing or are at a certain level of hearing that makes them unable to communicate effectively when using spoken language. Therefore, to help support the means of communication between the two communities, sign language was created. There are several types of communicating through media that can be used to interact with each other, one of which is using hand sign language communication media. The function of hand sign language is to help people with disabilities or what is commonly called disabled in communicating with each other. Hand sign language is also widely used by people who have behavioral disorders such as autism and Down syndrome. The commonly used hand sign languages are SIBI (Indonesian Signing System) and BISINDO (Indonesian Sign Language).

In order to participate in spreading the word about the existence of the original sign language from Indonesia, namely BISINDO, to all elements of society, a program is needed in the learning process that can correct when someone learns the BISINDO sign language. This program can be created by utilizing artificial intelligence technology or Artificial Intelligence (AI). The use of artificial intelligence uses one of its researches called machine learning, which allows computer programs to study data without interference from humans in making accurate predictions as if the computer program has intelligence [7]. The algorithm used in machine learning to create a sign language learning model is the Convolutional Neural Network (CNN), because it shows significant performance in handling data in the form of images [8]. With the CNN algorithm, computer programs can solve image recognition and classification problems so that they are suitable for recognizing sign language images [9]. However, between SIBI and BISINDO sign languages, SIBI has more and more diverse datasets available than BISINDO. This is because America uses the American Sign Language (ASL), which was adopted by SIBI, so that researchers especially in America develop ASL. Therefore, augmentation is needed on the BISINDO dataset, which aims to increase data.

Many studies have been carried out on this issue. In 2017 Indra et al conducted a study on BISINDO sign language using Chain Code Contour and Euclidean Distance tiwh static hand gestures recognition [10]. Indra et al in 2019 then continued to proposed a method to recognize BISINDO letters based on hand-shape features by using calculation difference in distance between query shape feature to each shape feature in database [11]. The other sign language have been studied by Nurul et al on words gestures in SIBI sign language using Kinect in 2017 [12]. Asriani et al in 2010 also researched on SIBI sign language gestures system using Backpropagation Neural network (BNN) with static hand gestures [13]. Research related to sign language, among others [14], [15], [16], [17], [18], and etc.

In this study, object localization and alphabets detection classification method are implemented to support intelligent systems in applications. The object localization method using Regional-Convolution Neural Network (R-CNN) is applied to mark the area of the finger circuit as a BISINDO signal in an image. While the CNN algorithm is used to provide decisions on the results of alphabetic detection classification which are recognized as BISINDO signals in the image of the localized object.

2 Methodology

In this section, we will discuss the methodology used in this research. The discussion starts from data pre-processing, architectural design, to system output.

2.1 Data Pre-processing

This research uses image data of BISINDO sign language. The number of data is 2080 consisting of 80 images for each alphabet class from A to Z. This data has been uploaded on the Kaggle website by researchers and can be downloaded for free at https://www.kaggle.com/idhamozi/indonesian-sign-language-bisindo. An example of BISINDO sign language image data is shown in Fig. 1.

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 143-149, October, 2023
ISSN: 2668-778X
www.techniumscience.com

Fig. 1. Image Data of BISINDO sign language.

This data is pre-processed before being used in the training and testing process. Data pre-processing for the classification and localization of objects is different in this study. Data pre-processing for object localization includes changing the image size to 128x128 and normalizing the image orientation. While data pre-processing for classification includes changing the image size to 128x128, converting the image to greyscale , and normalizing image orientation.

Augmentation is also needed to give variation to the data. In this study, data augmentation was only carried out for classification, but not for localization. Augmentation for training data on classification includes 6 processes, namely zoom range, width shift, height shift, horizontal flip, shear, and set fill mode. While augmentation for test data is only horizontal flip.

2.2 Architectural Design

The following is the system design for detection of BISINDO sign language in this study. Fig. 2 shows two core parts that dominant the system. The first part is the object localization process which aims to provide a bounding box for the sign language in the image. The second part is the classification process for alphabet detection in BISINDO sign language images.

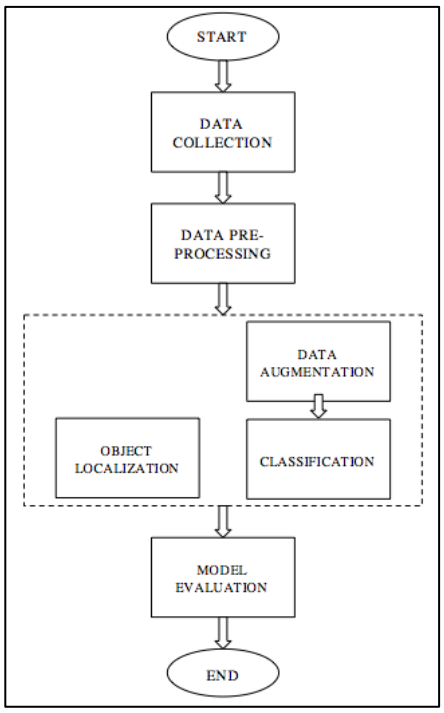


Fig. 2. System design

Two types of Neural Network algorithms are used in architectural design. The Region Convolution Neural Network (RCNN) algorithm is used to process object localization and the CNN algorithm will perform classification process.

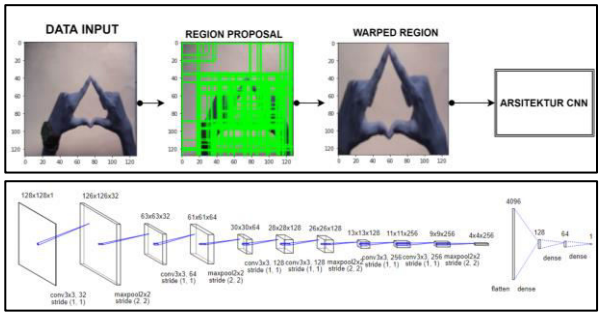


Fig. 3. The architectural design of object localization

Fig. 3 is the R-CNN architecture for localizing hand sign language objects. Object localization detects whether there is a hand object in the input image or not. Initially the 128 × 128 size input image is processed along with hand annotations. Then the proposal region is searched using selective search. Each region result will be calculated intersect over union then the image is cropped and entered into the CNN architecture to learn whether it is hand or not.

The CNN architecture for object localization learning is divided into two sequential parts, namely feature extraction and classification. The layers in the feature extraction are composed of six convolution layers and four pooling layers. Each convolution layer is equipped with dropout functions, batch normalization and ReLU activation functions.

The next process after feature extraction is classification. In this process the image data enters the fully connected layer. The first process in this section is flattening. Flattening serves to change the shape of a

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 143-149, October, 2023
ISSN: 2668-778X
www.techniumscience.com

multidimensional array into a one-dimensional array. In the fully connected layer, there are 2 hidden layers with 128 and 64 neurons. Then the outer layer has 1 class, namely hands. Because the outer layer contains only one class, the activation function used is sigmoid and the loss function used is binary crossentropy.

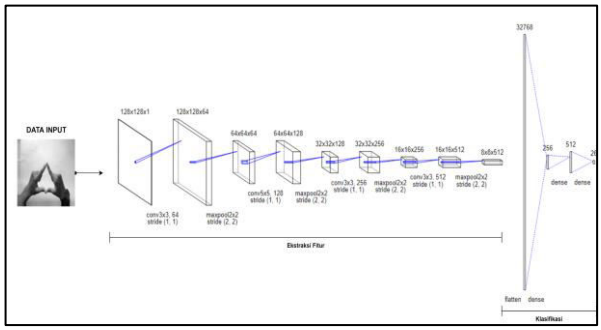


Fig. 4. The architectural design of classification

Fig. 4 is the CNN architecture for alphabetical classification A-Z in BISINDO sign language images. The layers in the feature extraction are composed of four pairs of convolution and pooling layers which are equipped with dropout functions, batch normalization and ReLU activation functions for each convolution layer. Initially the image has a size of 128x128x1 according to the results of pre-processing data. Then the extraction process is carried out on the convolution and pooling layer. This process is continued until the last layer.

After feature extraction process is finished, the image data enters the fully connected layer. The fully connected layer has two hidden layers with a sum of 256 and 512 neurons respectively. These two hidden layers are also equipped with dropout, batch normalization, and ReLU activation functions. The last layer after the hidden layer is the outer layer. This last layer consists of 26 classes, which are 26 BISINDO sign language alphabets from A to Z. The activation function used is softmax and the loss function used is sparse categorical crossentropy.

2.3 System Output

As explained in the previous section, this research is divided into two parts, namely object localization and classification. This means that there are two output systems in this study. The first output system is marking the bounding box of hand objects in sign language images. The second output system is the result of the alphabet A-Z classification in the BISINDO sign language image.

The training and testing process is required to obtain and measure the results of these two system outputs. The training process for each system uses 80% of the total data. There are 1664 data, consisting of 64 data in each alphabet class A-Z. The remaining 20% of the total data is used for the testing process. There are 416 data grouped into 26 alphabetic classes with the same proportions.

3 Result and Discussion

The following discusses the experimental results of the architectural design implementation. This experiment includes the stages of object localization and classification. In each experiment, 1664 data were trained in 50 epochs and 416 data were tested to calculate the confusion matrix and its accuracy.

3.1 Object Localization Experiments

The process stages that are passed in the object localization experiment start from the data generation process, the training process, and then the testing process. In the data generation process, the training data will be pre-processed first and then annotated to get the object coordinate label as a reference for detection accuracy.

The data pre-processing process is described in the previous section. It includes changing the image size to 128x128 and normalizing the image orientation. While the annotation process of image data is described in Fig. 5. This image annotation has been carried out one by one from the data. It aims to get the x, y, width and height coordinates as ground truth in determining how accurate the predicted bounding box is.

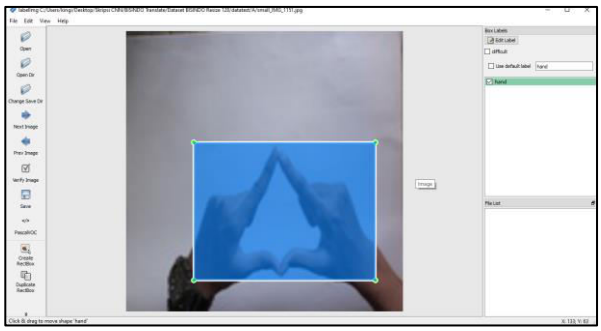


Fig. 5. The Image annotation process

After all the data has been generated, then the data will be processed in training. The learning rate used in this training is 0.00005. The training process will produce a model. It will be tested to measure the accuracy of the system architecture performance.

There are two types of tests carried out in this study. These are the testing using the training data itself and the testing using the actual test data. The testing using the training data is used to assess whether the model obtained is quite ideal. In this study, the experimental results in this first testing, in 100% accuracy. It means that all the training data successfully obtained the position of the bounding box well.

Fig. 6 is an example of the object localization testing results using one of the sign language image data. In the picture, it can be seen that the bounding box has successfully marked the hand object.

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 143-149, October, 2023
ISSN: 2668-778X
www.techniumscience.com

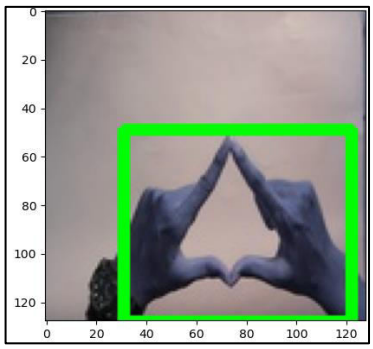


Fig. 6. The testing process of object localization

The other testing used the actual test data. This testing data certainly has never done training before. The results of this experiment are presented in Table 1. It contains the value of the confusion matrix from the testing result for each image in each A-Z alphabet class. Based on the information presented in the table, the accuracy of the testing results is 90,10%.

Table 1. Confusion matrix of object localization

Data	Presisi	Recall	F1-Score
A	1,00	1,00	1,00
B	1,00	1,00	1,00
C	1,00	1,00	1,00
D	1,00	1,00	1,00
E	1,00	1,00	1,00
F	1,00	1,00	1,00
G	1,00	1,00	1,00
H	1,00	1,00	1,00
I	1,00	1,00	1,00
J	0.04	0.01	0.02
K	1,00	1,00	1,00
L	0.99	1.00	0.99
M	1,00	1,00	1,00
N	1,00	1,00	1,00
O	1,00	1,00	1,00
P	1,00	1,00	1,00
Q	1,00	1,00	1,00
R	0.98	1.00	0.99
S	1,00	1,00	1,00
T	0,96	1,00	0,98
U	1,00	1,00	1,00
V	1,00	1,00	1,00
W	1,00	1,00	1,00
X	1,00	1,00	1,00
Y	1,00	1,00	1,00
Z	1,00	1,00	1,00
Average Value	**0.90**	**0.90**	**0.90**

The error of this testing occurs because there are some image data that the bounding box fails to detect object. Some of the image data are included in the alphabetical class J, L, R, and T. Although not all data in this alphabetic group has failed detection, this needs to be observed. Observations were made on the shape and position of the palm object in the image. The results of the analysis show that the data which the object localization fails to detect is caused by the distance of the hand object being taken too far from the camera. It is causing the object size in the image to be small.

3.2 Classification Process

The second experiment is the classification process. The purpose of this process is to detect sign language images into alphabet classes A-Z. Similar to the object localization process, in this experiment the data preparation, training process, and testing process were also carried out. The data preparation stage here only pre-processes the data without the annotation process. As explained earlier, that the data pre-processing stage in the classification process includes 3 processes, namely changing the image size to 128x128, converting the image to greyscale , and normalizing image orientation.

Fig. 7 shows an example of the data pre-processing result. The image on the left is the original image data, while the image on the right is image data after pre-processing.

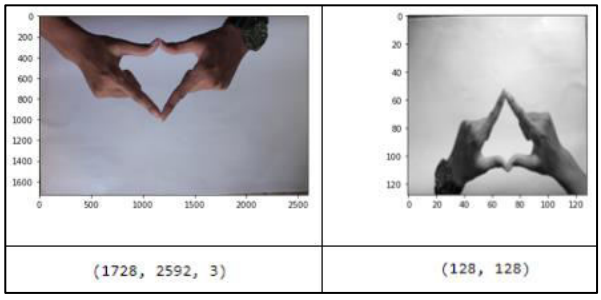


Fig. 7. The example of data pre-processing result

After the data was pre-processed, then the data will be processed in training. The learning rate used in this training is 0.00001. In **Fig. 8** is a visualization of changes the value of loss and accuracy during the training and validation process. The left graph is a visualization of the change the loss value, which is on the y-axis and the epoch on the x-axis. While the right graphic a visualization of changes in the accuracy value which is on the y-axis and the epoch on the x-axis. The orange graph data shows the experimental results of the validation data, while the blue ones are the results of experiments with training data. In the picture it can be seen that the graph moves down for the loss value while moving up for the accuracy value.

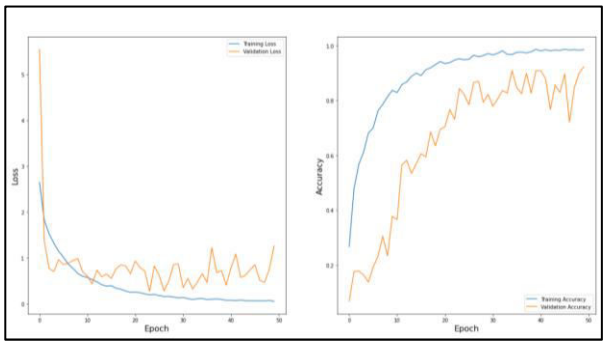


Fig. 8. The visualization of metric

Same as test scenario for object localization, there are two types carried out in this study. The first testing uses the training data itself and the results of this experiment is 100% accuracy.

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 143-149, October, 2023
ISSN: 2668-778X
www.techniumscience.com

Fig. 9 is an example of the classification test results. It can be seen in the figure, one image data of the Y alphabet sign language passed the testing process. The test results show that the sign language palm object is detected as the Y alphabet.

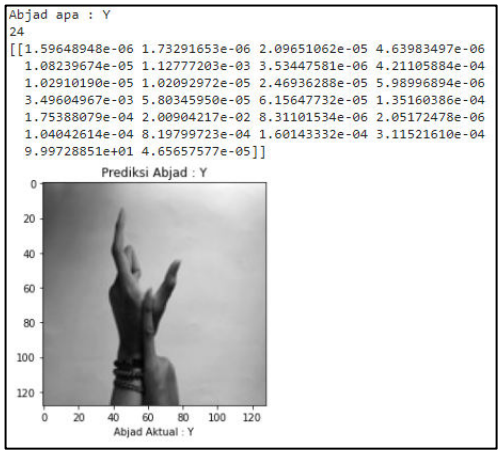


Fig. 9. The example of classification test result

The other testing using the actual test data are also carried out. The results of this experiment are presented in Table 2. Based on calculations from the data in the table, the accuracy of the testing results is 97,33%.

Table 2. Confusion matrix of sign language classification

Data	Presisi	Recall	F1-Score
A	1,00	1,00	1,00
B	1,00	1,00	1,00
C	1,00	1,00	1,00
D	1,00	1,00	1,00
E	1,00	1,00	1,00
F	1,00	1,00	1,00
G	1,00	1,00	1,00
H	1,00	1,00	1,00
I	1.00	0.69	0.81
J	0.94	1.00	0.97
K	1,00	1,00	1,00
L	1.00	0.94	0.97
M	1,00	1,00	1,00
N	1,00	1,00	1,00
O	0.94	1.00	0.97
P	1,00	1,00	1,00
Q	1,00	1,00	1,00
R	0.67	1.00	0.80
S	1,00	1,00	1,00
T	0,96	1,00	0,98
U	1.00	0.88	0.93
V	0.94	1.00	0.97
W	1,00	1,00	1,00
X	1,00	1,00	1,00
Y	1,00	1,00	1,00
Z	1.00	0.81	0.90
Average Value	**0.98**	**0.97**	**0.97**

At this testing stage, the classification was successful, although there were still a number of letters whose classification results were incorrect, such as the alphabets I, J, L, O, R, T, U, V, and Z. This happened because there were letters that were similar to each other, for example, the alphabet I was detected as the alphabet R.

4 Conclusion

Based on the results of the experiments that have been carried out, it can be concluded in general that the implementation of the design architecture in the object localization and classification stages has been able to detect hand sign language images, although not one hundred per cent.

The system has been able to detect and classify images from the BISINDO sign language alphabet, which produces an accuracy of 90.10% and 97.33% respectively.

There are several suggestions that allow it to be further developed on similar systems as follows: (1) The amount of data used should be reproduced. For the author's image data, using sign language data taken at almost the same time so that the lighting is similar, it would be better if the data used to further develop is sign language images that have a variety of lighting and a variety of skin colors; (2) Changing hyperparameters in the training process can be replaced with other values, some hyperparameters can be changed such as learning rate, batch or epoch. (3) Doing development on different architectures can enable to achieve better results.

References

1. Soekanto, Soerjono dan Budi Sulistyowati.(2014). Sosiologi Suatu Pengantar.Jakarta: PT. RajaGrafindo Persada.
2. Haryanto, Sidung. (2013). Spektrum Teori Sosial: Dari Klasik Hingga Postmodern. Yogyakarta: Ar-Ruzz Media
3. Turner, Lynn, and Richard West. 2013. Looseleaf for Introducing Communication Theory: Analysis and Application. McGraw-Hill Education.
4. Aththar, Muhammad Ahmad. 2012. The Magic of Communication. Serambi Ilmu Semesta.
5. Effendy, Onong Uchjana. 2005. Ilmu Komunikasi, Teori dan Praktek. Bandung: PT. Remaja Rosdakarya.
6. Alo Liliweri. 2010. Komunikasi Antar Pribadi, Bandung: PT. Citra Aditya Bakti
7. Ehsan Othman, A. A.-H. (2018). Automatic arabic document classification based on the hrwitd algorithm. Journal of Software Engineering and Applications.
8. Simonyan, Karen, dan Andrew Z., 2014 "Very deep convolutional networks for large scale image recognition." arXiv preprint 1409.1556
9. Bheda, V. and Radpour, D. (2017). Using deep convolutional networks for gesture recognition in american sign language. CoRR, abs/1710.06836.
10. D. Indra, S. Madenda, and E. P. Wibowo, “Recognition of Bisindo alphabets based on chain

code contour and similarity of Euclidean distance," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 5, pp. 1644–1652, 2017, doi: 10.18517/ijaseit.7.5.2746.

11. D. Indra, Purnawansyah, S. Madenda, and E. P. Wibowo, "Indonesian sign language recognition based on shape of hand gesture," Procedia Comput. Sci., vol. 161, pp. 74–81, 2019, doi: 10.1016/j.procs.2019.11.101.
12. W. Nurul Khotimah, Y. A. Susanto, and N. Suciati, "Combining decision tree and back propagation genetic algorithm neural network for recognizing word gestures in Indonesian Sign Language using Kinect," vol. 95, pp. 292–298, 2017.
13. F. Asriani and H. Susilawati, "Pengenalan Isyarat Tangan Statis Pada Sistem Isyarat Bahasa Indonesia Berbasis Jaringan Syaraf Tiruan Perambatan Balik," MAKARA Technol. Ser., vol. 14, no. 2, 2011, doi: 10.7454/mst.v14i2.709.
14. S. Subburaj and S. Murugavalli, "Measurement : Sensors Survey on sign language recognition in context of vision-based and deep learning," Meas. Sensors, vol. 23, no. May, p. 100385, 2022, doi: 10.1016/j.measen.2022.100385.
15. A. Ardiansyah, B. Hitoyoshi, M. Halim, N. Hanafiah, and A. Wibisurya, "Systematic Literature Review: American Sign Language Translator," Procedia Comput. Sci., vol. 179, no. 2020, pp. 541–549, 2021, doi: 10.1016/j.procs.2021.01.038.
16. Y. Wang, M. Du, K. Yu, G. Shen, T. Deng, and R. Wang, "Acta Psychologica Bi-directional cross-language activation in Chinese Sign Language (CSL) -Chinese bimodal bilinguals," Acta Psychol. (Amst)., vol. 229, no. January, p. 103693, 2022, doi: 10.1016/j.actpsy.2022.103693.
17. I. A. Adeyanju, O. O. Bello, and M. A. Adegboye, "Machine learning methods for sign language recognition: A critical review and analysis," Intell. Syst. with Appl., vol. 12, p. 200056, 2021, doi: 10.1016/j.iswa.2021.200056.
18. A. Imran, A. Razzaq, I. A. Baig, A. Hussain, S. Shahid, and T. ur Rehman, "Dataset of Pakistan Sign Language and Automatic Recognition of Hand Configuration of Urdu Alphabet through Machine Learning," Data Br., vol. 36, p. 107021, 2021, doi: 10.1016/j.dib.2021.107021.