PAPER • **OPEN ACCESS**

Indonesian Sign Language Recognition using YOLO Method

To cite this article: Steve Daniels *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1077** 012029

View the article online for updates and enhancements.

The Electrochemical Society
Advancing solid state & electrochemical science & technology
240th ECS Meeting ORLANDO, FL
Orange County Convention Center Oct 10-14, 2021
Abstract submission due: April 9 SUBMIT NOW

This content was downloaded from IP address 158.140.167.87 on 02/04/2021 at 15:23

Indonesian Sign Language Recognition using YOLO Method

Steve Daniels, Nanik Suciati*, and Chastine Fathichah

Department of Informatics, Faculty of Intelligent Electical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

E-mail: *nanik@if.its.ac.id

Abstract. Sign language is a form of communication commonly used by people with hearing impairment or people with speech impediments. Not all ordinary people understand the language. The translation of sign language into the alphabet/text automatically will facilitate the communication of the deaf with ordinary people. This research aims to develop a sign language recognition system that can process input from video data using You Only Look Once (YOLO) in real-time. YOLO is an object detection method based on Convolutional Neural Network (CNN), which is accurate and fast. Retraining the Yolov3 pre-trained model is performed with adjustments to the number of channels and classes according to the sign language recognition requirement. In this research, we collect datasets independently based on the Indonesian Sign Language (BISINDO). In the experiment using image data, the system achieves 100% precision, recall, accuracy, and F1 score. While using video data, the system's performance gets precision 77.14%, recall 93.1%, accuracy 72.97%, and F1 score 84.38%, with a speed of 8 fps.

1. Introduction

Sign language is a form of non-verbal human communication expressed through hand movements to convey information to the recipient visually. Sign language is commonly used by people with hearing impairment or people with speech impediments. Most normal people do not learn sign language because it is not required to do. But if there's chance of a need to understand sign language, people cannot just be able to interpret sign language instantly without learning process. Therefore, a system that automatically recognizes and translates the sign from images or videos to alphabet/text will provide a way to understand sign language and ease communication between deaf/mute and ordinary people.

There are several previous researches that have done in sign language recognition. The type of approaches can be divided into two categories. The first approach is special equipment based, such as accelerometer [1], sensory glove [2], and Kinect [3]. In special equipment based approach usually there are two work steps. The step are capturing sensory signal from sensory equipment and recognizing sign from captured signal. This approach can provide accurate information, but there is a need for using particular equipment which can be either expensive or impractical for daily use.

The second approach is vision based. This approach is focused on vision features, which is lower cost in terms of equipment, requires only a camera for taking images or videos, and highly practical for daily usage because we can use webcam or smartphone camera as well. For example, there is a research using kNN classifier to recognize sign [4]. From computer vision and machine learning techniques combined, Convolutional Neural Network (CNN) was proposed and has become common method for vision related machine learning and been applied to various task such as object detection and recognition.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

The CNN method was once used to translate American Sign Language (ASL) into the alphabet [5]. The CNN architecture used in that research is GoogLeNet, which has been trained on the 2012 ILSVRC dataset. The experiment obtained an accuracy of 72%. This lack of accuracy is due to the misclassification of alphabets with similar handshapes, for example, g / h and m / n / s / t.

One of the deep learning methods that have recently been widely used for object detection is "You Only Look Once" (YOLO). YOLO is developed based on Convolutional Neural Network (CNN) and can produce fast and effective object detection [6]. YOLO is used in various studies that require object detection, such as real-time localization of Bhutan number plates [7], pedestrian detection [8], and recognition of traffic signs [9].

In this research, we implement CNN with YOLO architecture to recognize Indonesian Sign Language (BISINDO) from video data. Transfer learning is used to retrain the YOLOv3 pre-trained model with adjustments to the number of channels and classes according to the Indonesian Sign Language requirement. We collect a dataset independently consists of 4,547 images with 24 classes of static sign language. We use 80% of the dataset for training 20% for testing. Performance targets measured include precision, recall, accuracy, F1 score, mAP, and fps.

This article is organized into five chapters. Chapter 1 contains the background and research objectives; Chapter 2 explains the theoretical basis; Chapter 3 provides an overview of the system design; Chapter 4 describes the results of the experiment and analysis, and Chapter 5 represents conclusions.

2. Theory

2.1. Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of neural network that is commonly used in image data. CNN consists of neurons that have weight, bias, and activation functions. Generally, CNN consists of two parts, namely feature extraction layer and fully connected layer.

Feature extraction layer consists of 2 layers, convolutional layer and pooling layer. Convolutional layer consists of neurons that are arranged to form filters. This filter is shifted to all parts of the image, a dot operation is performed between the input and the value on the filter. The result of the process at this layer is a feature map. The feature map from the convolutional layer then passes through the pooling layer, a layer consisting of a filter with a certain size and stride which is shifted to all parts of the feature map. The purpose of the pooling layer is downsampling, which is to reduce the dimensions of the feature map. The feature map generated from the feature extraction layer is still in the form of a multidimensional array. Therefore, flattening is performed to change the shape of the feature map into a vector to be used as input for the fully connected layer. The output of this layer is the output of the network.

2.2. You Only Look Once

You Only Look Once (YOLO) is an object detection method known for its speed. Only a single convolutional network used to predict what objects will be in the image. This can happen because YOLO divides the image into cells/grids, each cell is responsible for predicting the number of bounding boxes, the confidence level of each cell, and the class probability.

2.3. YOLOv3

YOLOv3 is an improvement of the basic idea of YOLO, dividing an image into cells that are responsible for predicting objects. Changes from YOLO include feature extraction networks, use of bounding boxes, and detection at various scales [10]. YOLOv3 architecture consists of a feature extraction network and a detection layer with three different scales, which can be seen in Figure 1.

The feature extraction network used only consists of a Convolutional Layer, followed by batch normalization and the Leaky ReLU Activation function. Batch normalization is useful for helping the convergence model process during training. This network also has a residual block that passes results

from one layer to the next layer and a deeper layer directly, so there is no degradation problem in a network that has many layers.

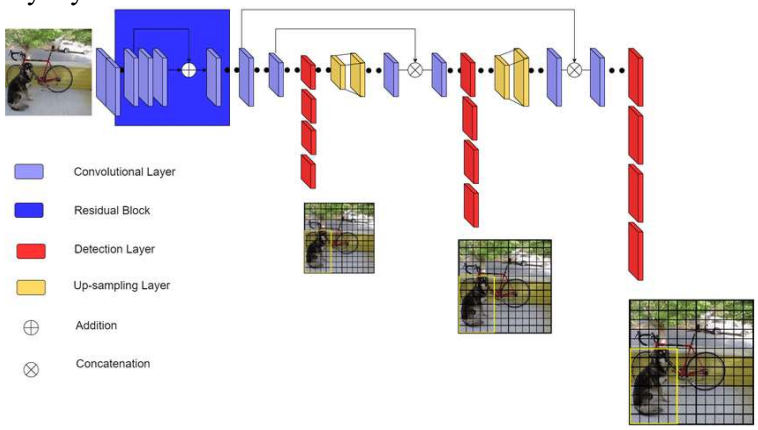

Figure 1. YOLOv3 Architecture [11]

The use of anchor boxes is done to replace predicting width and height of the bounding box directly because predicting the width and height of the bounding box causes an unstable gradient during training. Bounding box prediction is done by transforming the anchor box that has been assigned from the start. The object score and class score will enter the sigmoid function so that the results are between 0 to 1.

YOLOv3 performs predictions on three different scales. The detection layer is used for three feature maps of different sizes, each having a stride of 32, 16, and 8. With an input of 416×416, detection is carried out at a scale of 13×13, 26×26, and 52×52.

2.4. Non Maximum Suppression

Non-Maximum Suppression (NMS) is used by YOLO to eliminate bounding boxes that predict the same object. The way it works is to take the bounding box with the highest confidence then eliminate other bounding boxes that make predictions in the same class and have a similar location/overlap that exceeds the specified threshold.

2.5. Bahasa Isyarat Indonesia

Bahasa Isyarat Indonesia (BISINDO) is one of the sign languages used in Indonesia and is a language that grows naturally among the Deaf community in Indonesia. BISINDO alphabet signs are in the form of a static hand, except for the letters J and R which use gestures. BISINDO sign can be seen in Figure 2.

Figure 2. BISINDO sign [12]

3. Design

The design of the system to recognize Indonesian sign language is shown in Figure 3. The training process is performed to get a classifier model used later in the testing process to recognize sign language on image or video.

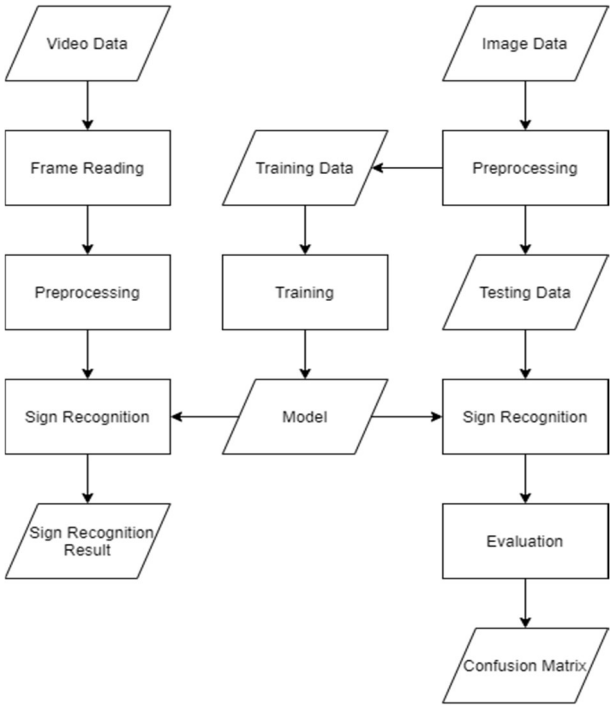

Figure 3. Diagram of the Indonesian sign language recognition

3.1. Dataset

The data used as input for the sign language recognition system is a dataset taken by the author in the form of photos and videos, consisting of 24 classes where each class there are about 160 to 220 images. This data is a set of images of characters, A to Z except J and R. Dataset specification is given in Table 1.

Table 1. Dataset Specification

Specification	Value
Resolution	3024 × 3024, 640 × 480
Extension	.jpg
Number of images	4,547
Number of class	24
Number of images per class	160-220
File size	50-500 kB
Channel	3 (RGB)

The dataset is split into training and testing data with a ratio of 8: 2; 8 for training data and 2 for trial data. The dataset is taken with different subjects and environments, a total of 5 conditions which are described in Table 2, and sample of each condition can be seen in Figure 4.

Table 2. Dataset Condition

Condition	Description
Condition 1	Slightly dim, taken from 20-40 cm away
Condition 2	Bright, taken from 20-40 cm away
Condition 3	Bright, taken from 20-80 cm away
Condition 4	Bright, taken from 20-40 cm away
Condition 5	Bright, taken from 150 cm away

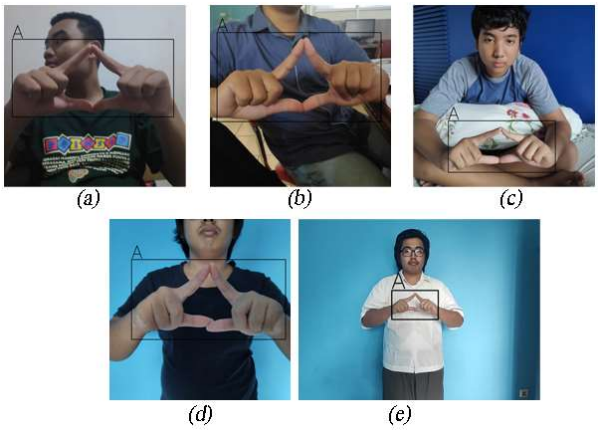

Figure 4. Sample of: (a) Condition 1; (b) Condition 2; (c) Condition 3; (d) Condition 4; (e) Condition 5

3.2. Preprocessing

At this stage, data preprocessing will be carried out before becoming input to the training process. The preprocessing data includes resizing to 416 × 416, conversion from RGB to grayscale, and adding data by rotating the available images by 1 and 2 degrees clockwise and counterclockwise. Preprocessing is done so that the data varies and equates the size of the training data. The result example can be seen in Figure 5.

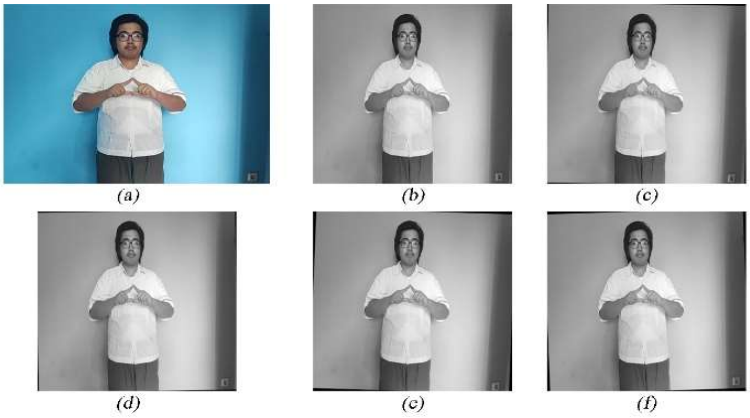

Figure 5. (a) Original image; (b) After resizing and converting to grayscale; (c) and (d) After rotating 1° counterclockwise (CCW) and clockwise (CW); (e) and (f) After rotating 2° CCW and CW.

3.3. Training

The training stage uses 80% of preprocessed dataset, while the rest is used for testing. The split is made so that train and test dataset has portions of all conditions. Adjustments are made from the original YOLO architecture, where changes occur in the number of channels and the number of filters in the convolutional layer before entering the Yolo layer. The training process utilizes training data to build a YOLO model. Training is done by using pretrained weight darknet53, which was trained on ImageNet, with various learning rates. Training will be stopped when the loss value does not decreasing anymore.

3.4. Testing

This phase has two types of testing, namely, testing of image data and video data. In image testing, several images are detected. At the end of the testing, a comparison of the detection results with the ground truth will be carried out, creating a Confusion Matrix to obtain accuracy, precision, recall, and

F1 score. Mean Average Precision (mAP) is also used to evaluate object detection task. In video testing, detection will be performed on each frame of the video data. If a class is detected sequentially more than the specified limit, it will be considered a recognized sign. The results of the recognized sign will be compared with the ground truth.

4. Experiments

In this chapter, there are several experiment scenarios: training results, image testing results, and video testing results.

4.1. Training Results

The training was carried out by forming three models with different learning rate parameters, namely 0.1, 0.01, and 0.001. The model is trained until the loss value does not decrease. Each model is compared based on the lowest loss value obtained. The training results can be seen in Table 3.

Table 3. Training Results

Learning Rate	Loss
0.1	0.942662
0.01	0.15379
0.001	0.143697

From Table 3, the model trained using a learning rate of 0.001 achieves the lowest loss value. This model will be used for subsequent test scenarios.

4.2. Image Testing Results

Testing on image data is done by running recognition on all images from the testing data. Some of result images can be seen in Figure 6.

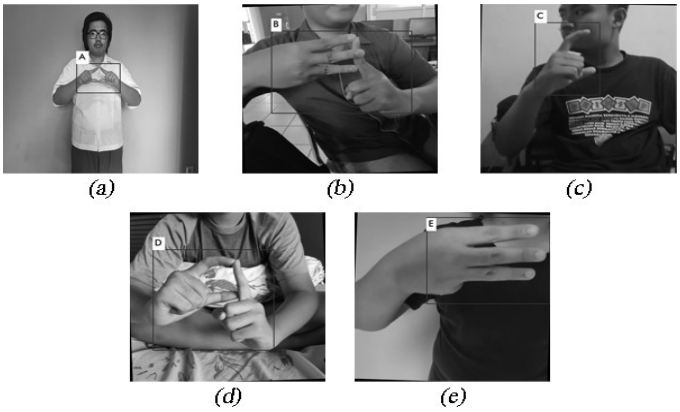

Figure 6. Prediction result examples: (a) A Sign, (b) B Sign, (c) C Sign, (d) D Sign, and (e) E Sign

Table 4. Object Detection evaluation

IoU Threshold	mAP (%)
0.75	71.16
0.5	99.91
0.25	99.96

Object detection task is tested with mAP as the evaluation metric. Three different Intersection over Union (IoU) threshold values is being used for this test. The test results can be seen in Table 4. The prediction results will be compared with the ground truth provided by the author. Detection is limited to

1 detection with the highest confidence for each image. The assessment will be carried out using precision, recall, accuracy, and F1 score. The test results can be seen in Table 5.

Table 5. Image Testing Results

Threshold (%)	Precision (%)	Recall (%)	Accuracy (%)	F1 Score (%)
100	100	41.67	42.20	58.83
90	100	76.53	76.97	86.71
80	100	83	83.33	90.71
70	100	87.72	87.90	93.46
60	100	91.56	91.62	95.59
50	100	94.53	94.57	97.19
40	100	96.70	96.72	98.32
30	100	98.27	98.28	99.13
20	100	99.51	99.52	99.75
10	100	99.98	99.98	99.99
9	100	100	100	100

As shown in Table 5, all the detections carried out resulted in the correct class. However, some images have low confidence so that they are not detected at a certain threshold.

Class O has the highest number of detections with confidence value below 80%. The detection result that has the smallest confidence is an image of O sign with a confidence of 9%.

Apart from being based on class, observations were also made based on image conditions. The image condition that has the most amount of low confidence is condition 5, where the cue image is the farthest away. From 247 cases of undetectable images at the threshold of 50%, 246 cases came from images with condition 5.

Table 6. Video Testing Results

Threshold (%)	Target	Result	Explanation
6	GLEAM	GLEAM	-
	IVY	OIOY	Prediction error of I becomes O, and V becomes O
	NICE	MICE	Prediction error of N becomes M
	SHIZA	SLZA	Alternating prediction of H between D and H without sequential detection of 5, the prediction error I becomes L
	WAX	WADA	Prediction error X becomes D, then becomes A
	BDFPUV	BPDBPDUVUV	The prediction error of D becomes P, F becomes B, P becomes D, U becomes V
	QUEST	QUEIST	Detection of I reached 6 sequentially when the transition of signals between E to S
7	WAX	WA	Prediction error of X becomes A
	BDFPUV	BPDBPDVUV	The prediction error of D becomes P, F becomes B, P becomes D, U becomes V
	QUEST	QUEST	Number of detected I sequentially is below 7

4.3. Video Testing Results

Testing on video data was carried out on several videos taken by the author, separated from the data taken by training in conditions similar to condition 5 (taken from 150 cm away). Testing videos have the same dimensions, namely 640 × 480. The video consists of two or more signs that are displayed consecutively, forming a series of letters according to the target letters of the video.

The number of objects that will come out as detection is limited to 1 with the highest confidence. The detection of signs in a sequence reaching the specified limit will be considered as a recognized sign and will be compared with the ground truth provided.

FPS calculation is done by calculating the time needed to perform signal recognition in one frame, then counting how many frames can be recognized in one second.

The results of testing on video data show a precision value of 80.56%, a recall of 93.1%, an accuracy of 72.97%, and an F1 score of 84.38%. Details of the experiment results is given in Table 6.

From the test results shown in Table 6, two errors were found. The first error is a class prediction error, for example, the letter 'N' in the 'NICE' target is known as' M ', also the' H 'in the' SHIZA 'target is not known because the prediction results obtained keep changing between' H 'and' D 'without successive predictions of reaching the specified streak limit. The second error is, sign recognition which occurs during the transition from one sign to another, for example, the prediction of the sign 'I' on the target 'QUEST' during the transition between signs 'E' to 'S' for 6 consecutive times.

The prediction error occurs is estimated for class N and M because they have similar terms, as well as P with D, U with V, I with L at a distance, and W with A at a distance.

For video detection, the speed ranges from 8-9 FPS for video input with dimensions of 640 × 480. 8 FPS speed cannot process all live input frames, which are generally 20-30 frames per second, in real-time. Therefore, for real-time usage, it is necessary to select a frame that will be recognized so that the frame recognized by the system is the newest frame that has just been obtained from the input. For example, if the system has a speed of 8 FPS and an input of 24 FPS, the frames to be detected are multiples of 3. With 8 frames detected every second and streak limit of 7, real-time sign language recognition can be done with sign language approaching a natural speed whereby the sign is carried out for about 1 second for each letter.

5. Conclusion

In this research, a recognition system for The Indonesian Sign Language using YOLO has been implemented. The experiment on image and video data got 100% and 72.97% accuracy, respectively. Recognition of the transition frames between one sign to another on video data contributes to the misrecognition error. An algorithm to distinguish a transition frame and a sign frame should be considered in the future to improve the accuracy of the system. With the processing speed of 8 fps, real-time recognition of video data can be performed by adjusting the speed of the recognition process and frames rate.

References

[1] Zafrulla Z, Brashear H, Yin P, Presti P, Starner T and Hamilton H 2010 American sign language phrase verification in an educational game for deaf children *2010 20th Int. Conf. on Pattern Recognition* pp 3846–9

[2] Oz C and Leu M C 2011 American Sign Language word recognition with a sensory glove using artificial neural networks *Eng. Applications of Artificial Intelligence* **24** pp 1204–13

[3] Lang S, Block M and Rojas R 2012 Sign language recognition using kinect *Int. Conf. on Artificial Intelligence and Soft Computing* pp 394–402

[4] Aryanie D and Heryadi Y 2015 American sign language-based finger-spelling recognition using k-Nearest Neighbors classifier *2015 3rd Int. Conf. on Information and Communication Technol. (ICoICT)* pp 533–6

[5] Garcia B and Viesca S A 2016 Real-time American sign language recognition with convolutional neural networks *Convolutional Neural Networks for Visual Recognition* **2** pp 225–32

[6] Redmon J, Divvala S, Girshick R and Farhadi A 2016 You only look once: Unified, real-time object detection *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 779–88

[7] Chen R C 2019 Automatic license plate recognition via sliding-window darknet-YOLO deep learning *Image and Vision Computing* **87** pp 47–56

[8] Qu H, Yuan T, Sheng Z and Zhang Y 2018 A Pedestrian detection method based on YOLOv3 model and image enhanced by Retinex *2018 11th Int. Congress on Image and Signal Processing, BioMedical Eng. and Informatics (CISP-BMEI)* pp 1–5

[9] Zaki P S, William M M, Soliman B K, Alexsan K G, Khalil K and El-Moursy M 2020 Traffic signs detection and recognition system using deep learning *Preprint arXiv:2003.03256*
[10] Redmon J and Farhadi A 2018 Yolov3: An incremental improvement *Preprint arXiv:1804.02767*
[11] Valdez P 2020 Apple defect detection using deep learning based object detection for better post harvest handling *Preprint arXiv:2005.06089*
[12] Peduli Kasih ABK (2018) *Mengenal Bahasa Isyarat* Retrieved on 18 December 2019 from https://www.ypedulikasihabk.org/2018/11/09/mengenal-bahasa-isyarat/