Recognizing Indonesian sign language (Bisindo) gesture in complex backgrounds

Muhammad Alfhi Saputra, Erdefi Rakun
Faculty of Computer Science, University of Indonesia, Depok, Indonesia

Article Info

Article history:

Received Jan 2, 2024
Revised Aug 13, 2024
Accepted Aug 26, 2024

Keywords:

Bisindo
Faster RCNN
Object detection
Sign language recognition
YOLOV5

ABSTRACT

Sign language, particularly Indonesian sign language (Bisindo), is vital for deaf individuals, but learning it is challenging. This study aims to develop an automated Bisindo recognition system suitable for diverse backgrounds. Previous research focused on greenscreen backgrounds and struggled with natural or complex backgrounds. To address this problem, the study proposes using Faster region-based convolutional neural networks (RCNN) and YOLOv5 for hand and face detection, MobileNetV2 for feature extraction, and long short-term memory (LSTM) for classification. The system is also designed to focus on computational efficiency. YOLOv5 model achieves the best result with a sentence accuracy (SAcc) of 49.29% and a word error rate (WER) of 16.42%, with a computational time of 0.0188 seconds, surpassing the baseline model. Additionally, the system achieved a SacreBLEU score of 67.77%, demonstrating its effectiveness in Bisindo recognition across various backgrounds. This research improves accessibility for deaf individuals by advancing automated sign language recognition technology.

Corresponding Author:

Erdefi Rakun
Faculty of Computer Science, University of Indonesia
16424 Depok, West Java, Indonesia
Email: efi@cs.ui.ac.id

1. INTRODUCTION

Communication is one of the essential aspects of human life. However, some people with a condition such as hearing impairment or deafness have limitations in verbal communication. People with such conditions usually communicate using sign language. There are two sign languages for the Indonesian language: the Indonesian sign language system (SIBI) and Indonesian sign language (Bisindo). Research on Bisindo is still less extensive than that on SIBI; therefore, further development and research are needed.

Research on sign language recognition is diverse, with various datasets and models [1]. There are two common approaches to sign language recognition (SLR) research: vision-based and glove-based. Vision-based research is considered more natural and realistic because it utilizes information based on the natural environment [2]. The main difference between the two methods is that the vision-based uses image or video data. Meanwhile, the glove-based system uses sensors [3]. For the vision-based approach, several methods can be used, including principal component analysis (PCA) [4], hidden Markov model (HMM) [5], support vector machine (SVM) [6], histogram of oriented gradients (HoG) [7]. In addition to those methods, deep learning models are also can be used for these cases; the frequently used deep learning models are 3D convolutional neural network (CNN) [8], generative adversarial network (GAN) [9], variational autoencoder (VAE) [10], Faster region-based convolutional neural network (RCNN) [11], [12], you only look once (YOLO) [13], [14], which are used as hand-locating networks to detect hands before the classification process.

SLR research development for Indonesian sign language still needs to be improved, especially for Bisindo. An SLR study for Bisindo utilizes the YOLOv3 model but is still limited to recognizing only the alphabet [15]. A study on Bisindo that carried out word-level recognition was done, but it still uses skin color segmentation to extract hands and faces [16]. However, the study on Bisindo is still limited to datasets with greenscreen background only. SLR research for SIBI has seen more development than Bisindo. The study on SIBI recommends a feature extraction model, which is MobileNetV2. The research still only used skin color segmentation to extract hands and faces [17]. It was subsequently proven that the system performed well only for datasets with a green background but needed to be more suitable for complex or natural backgrounds [18]. Complex backgrounds are crucial in sign language recognition research because real-world sign language data often have a natural background, not necessarily a green screen. Therefore, addressing the issue of complex backgrounds becomes essential.

SLR research on Bisindo still needs to recognize words or sentences effectively in complex backgrounds. Therefore, this research aims to develop a Bisindo sign language recognition system that performs well in complex backgrounds. Object detection is required at the initial stage of the process to address the issue of complex backgrounds. The models used for detection are Faster RCNN and YOLOv5. These two models have different characteristics, as Faster RCNN is a two-stage detector [19], and YOLOv5 is a single-stage detector [20]. Hopefully, comparing and integrating these two models into the system framework will determine the best model for recognizing Bisindo sign language. This research will also apply a measurement metric that has not been previously used for this case, namely SacreBLEU [21]. SacreBLEU is claimed to be more suitable for evaluating language translators where the original language and the translated language have similar language characteristics.

2. PROPOSED METHOD

Table 1 contains studies related to this research. Generally, recognizing sign language using video or image-based methods involves detecting hands and faces, extracting features, and classifying sign language words. In the research conducted by He [11], a method was proposed for hand detection using Faster RCNN, which will be employed in this study to detect hands and faces. Padmaja *et al.* [22] also employed a similar method but could only detect several letters and one word, whereas this research aims to detect complete sentences. This study also utilizes YOLOv5 as a lighter model compared with Faster RCNN. The method for feature extraction will utilize the findings from Setyono and Rakun [17], which compared MobileNetV2 and ResNet50, concluding that MobileNetV2 is more efficient, especially for implementation on mobile devices. The classification method will implement the findings from Faisal [16], which tested various long short term memory (LSTM) [23] configurations, concluding that a 1-layer bidirectional LSTM (BiLSTM) is the most effective for Bisindo. For evaluation, in addition to using metrics that have been tested in previous studies, namely sentence accuracy (SAcc) and word error rate (WER), this study will also try other metrics, namely BLEU [24] and SacreBLEU [21], to determine which metric is most suitable for Bisindo sign language recognition research.

Table 1. Related works

Author	Specification	Remarks
He [11]	Proposing a hand locating network using Faster RCNN, followed by CNN and LSTM	This paper proposes a hand locating network to initially detect only hand objects for subsequent sign language recognition. The same principle is also applied in this research.
Padmadja *et al.* [22]	Building SLR using Faster RCNN and ResNet50	This paper shows that the Faster RCNN method can be used for sign language recognition, but the research is still limited to recognizing only a few letters and one word.
Setyono and Rakun [17]	Recognizing word gesture in sign system for Indonesian language (SIBI) Sentences using DeepCNN and BiLSTM	This paper compares the deep CNN methods as a feature extractor and obtains the MobileNetV2 model as the feature extractor. This model will be used in this research.
Daniels *et al.* [15]	Building SLR for recognizing Bisindo alphabet using YOLOv3.	This paper uses the Bisindo dataset, but it is limited to the alphabet only.
Faisal [16]	Development of Indonesian sign language movement recognition model (Bisindo) using Mobilenetv2 as feature extractor and LSTM as classifier	The study uses the Bisindo dataset and proposes a 1-layer bidirectional LSTM as the best classification method for Bisindo. This method will be used in this research.
Nimisha and Jacob [3]	Describing the SLR approaches in vision-based and glove-based methods. Comparing feature extraction and classification models.	Finding that YOLO is a good and fast model for sign language recognition. The advantage of YOLO in terms of very fast detection time will be tested in this research.
Dima and Ahmed [20]	Using YOLOv5 Algorithm to detect and recognize American sign language.	YOLOv5 successfully recognizes sign language with an mAP of 95%. This version of YOLO will be used in this study.

3. METHOD

The methods used in this study can be seen in Figure 1. In general, there are six stages, namely data preparation to prepare and preprocess data, object detection to detect hands and faces objects using Faster RCNN and YOLOv5 with the evaluation matrix of mean average precision-intersection over union (mAP-IoU), skin color segmentation to segment objects based on skin color using a multi-color space, feature extraction using MobileNetV2 and accuracy evaluation matrix, classification using LSTM and accuracy evaluation matrix, and finally, evaluation, which is the final measurement of the system with matrices including SAcc, WER, and execution time. Each stage is further elaborated in the following sections.

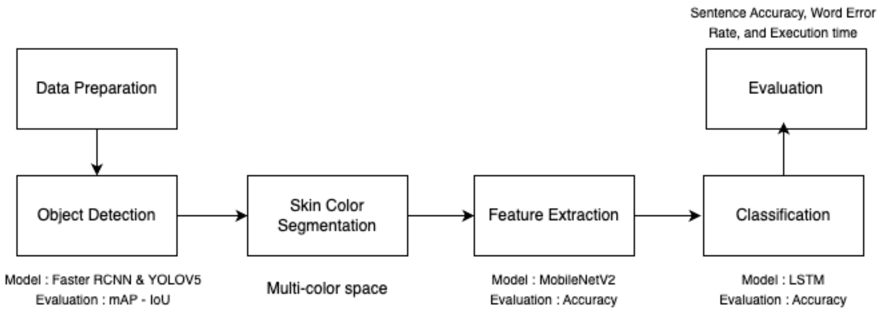

Figure 1. The overall process of experiment

3.1. Data preparation

The dataset used in this research consists of a collection of 420 videos demonstrating Bisindo. The device used to record those videos is a Samsung Galaxy S9+ with a resolution of 1,920×1,080 pixels and 30 frames per second. This dataset consists of four individuals, three performing 40 sign language sentences each and the fourth performing 20 sign language sentences. There are three videos for each sentence demonstrated by each individual. Look at Table 2 in Appendix for the details about the dataset.

Figure 2 illustrates the stages of data preparation. The first step in data preparation is converting the video dataset into a collection of frames. Next, the process includes cropping and resizing to obtain images with a 1:1 aspect ratio and a size of 224×224 pixels and then annotating the dataset with bounding boxes for the right-hand, left-hand, and face objects. After the data preparation, the total number of frames obtained is 92,977. The next step involves labeling the frame dataset with words for each frame. The dataset includes 152 words, each labeled from 1 to 152, with label 0 reserved for transitions.

In addition to the original dataset with a greenscreen background, a dataset with a complex background was also constructed. To create this complex background dataset, some images were applied to replace the greenscreen background in the initial dataset. The chosen images contain various natural objects with skin colors, forcing the system to learn more robustly. Figure 3 shows the images used as complex backgrounds for the complex dataset.

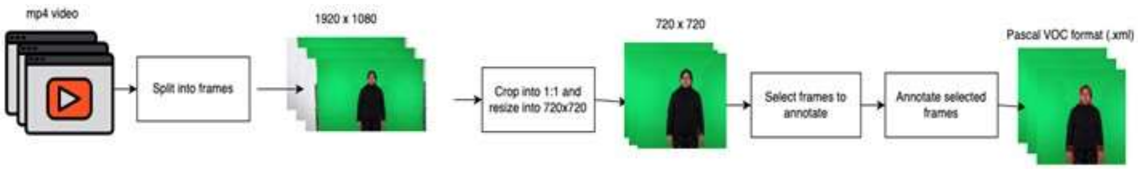

Figure 2. Data preparation

(1) (2) (3) (4) (5)

Figure 3. Images used as complex backgrounds

3.2. Object detection

Based on Figure 4, the object detection process involves taking the input dataset and annotations, resizing images to 224×224, splitting the data into training and testing sets with an 80%-20% proportion, training the Faster RCNN and YOLOv5 models, conducting testing, and evaluating using mAP-IoU. There are three object classes: the right hand, left hand, and face. The study that uses only skin color segmentation demonstrates that it can effectively extract hands and faces in datasets with a green screen background [18]. However, the method will also tested on complex datasets, and it would be the baseline method of this study.

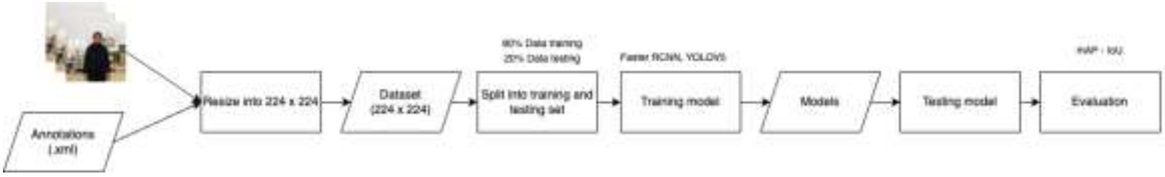

Figure 4. Object detection

3.3. Skin color segmentation

Rahmat *et al.* [25] introduced the skin segmentation method using multi-color space, which combines three color spaces: Normalized RGB, HSV, and YCbCr. The schema for skin color segmentation can be seen in Figure 5. Subsequently, Figure 6 displays the results of the segmentation. The figure consists of two subfigures: Figure 6(a) segmentation results without object detection and Figure 6(b) segmentation results with object detection. The difference in segmentation results on a complex background is apparent between the method only relying on skin color segmentation as shown in Figure 6(a) and the one containing object detection before skin color segmentation as shown in Figure 6(b). The method without object detection in Figure 6(a) still captures other objects in the background. In contrast, the use of object detection in Figure 6(b) ensures precise segmentation limited to the hands and faces of the performer, as intended.

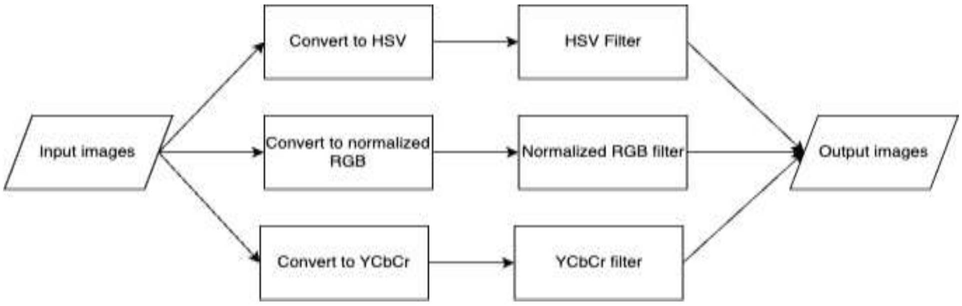

Figure 5. Skin color segmentation

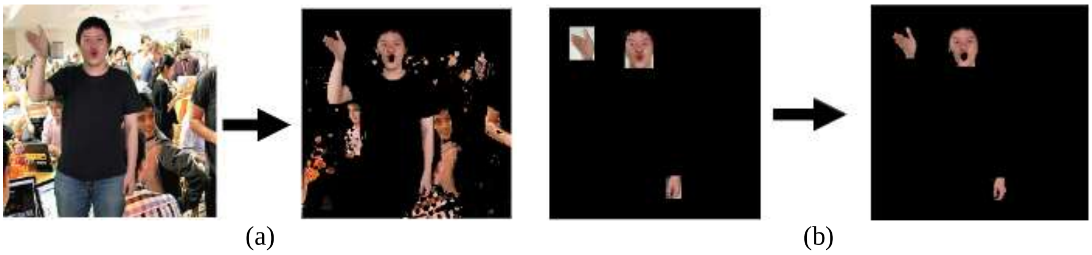
(a)
(b)

Figure 6. Comparison of segmentation results (a) without object detection and (b) with object detection

3.4. Feature extraction

Figure 7 displays how the feature extraction is performed. The feature extraction model learns features from the dataset, generates a model, and then applies it to the dataset for extraction. This study uses MobileNetV2 for feature extraction. It employs transfer learning, retraining the pre-trained MobileNetV2 model on this research dataset. This feature extraction process produces an output vector with a size of 1,280.

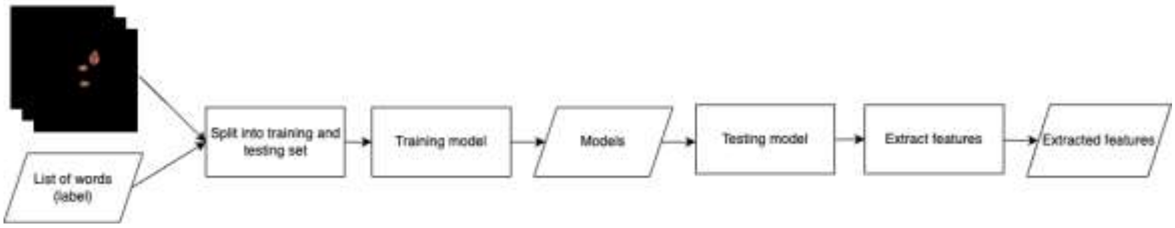

Figure 7. Features extraction

3.5. Classification

Based on Figure 8, the model used for classification is LSTM. The LSTM architecture used is a 1-layer BiLSTM, the best architecture for recognizing Bisindo sign language [16]. LSTM is a seq2seq model requiring the same input and output sequence length. The length of the input sequence, in this case, is ten. That means each label in each sentence will be represented with ten frames. The value ten is derived from the average length of frames for each label.

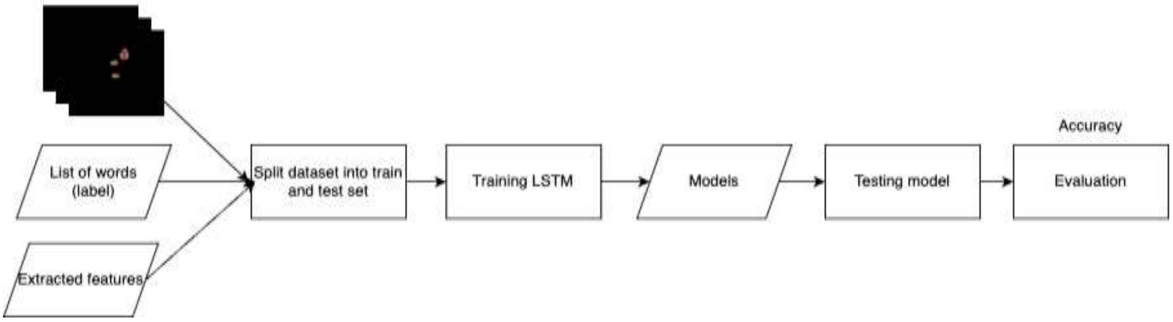

Figure 8. Classification

3.6. Evaluation

This research conducts a quantitative evaluation as follows: mAP to assess the performance of the object detection model by comparing ground truth with IoU area [26]; SAcc to measure the accuracy of the classified sentences; and WER to determine the level of word classification errors in assessing the classification results [27]. The formulas to calculate mAP, WER, and SAcc are:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{1}$$

$$WER = \frac{S+D+I}{N=H+S+D} \tag{2}$$

$$SAcc = \frac{Number\ of\ correct\ sentences}{Total\ number\ of\ sentences} \tag{3}$$

In this context, S (substitutions) represents the total number of substitutions, where each instance involves replacing a word in the sentence with a different one. D (deletions) represents the total number of deletions, referring to words that appear in the original sentence but are missing in the prediction. The variable I (insertions) represents the number of insertions, which are words not present in the original sentence but included in the prediction. C (correct words) indicates the count of words that are correct in the prediction. Finally, N (number of words) is the total number of words in the reference sentence, and it equals the sum of S, D, and C. A higher WER value indicates a worse system performance. Conversely, for SAcc, a higher value indicates better system performance. The overall system will be assessed by testing it on the entire test dataset, using the following formula to evaluate its performance. In addition to the metrics, the experiments calculate the processing time, which consists of the inference time or the time it takes for the system to process one input data and produce an output.

$$\mu WER = \frac{\sum WER}{n} \tag{4}$$

$$\mu SAcc = \frac{\sum SAcc}{n} \tag{5}$$

The other evaluation method used is SacreBLEU [21]. This evaluation method has yet to be used in previous research on Indonesian sign language. SacreBLEU measures how closely a system's translation or prediction approaches the reference or ground truth. However, SacreBLEU will be tried for sign language recognition in this case. Here is the formula.

$$BLEU = BP \, * \, exp(\textstyle\sum_{n=1}^{N} W_n . log(precision_n)) \tag{6}$$

3.7. Experiment design

This study evaluates the performance of a system with object detection versus one without, using both greenscreen and complex background datasets. Six experiments were conducted: the first two without object detection (serving as baselines) and the next four using Faster RCNN and YOLOv5 for hand and face detection across both datasets. Results were compared to assess improvement, with the ANOVA significance test used for evaluation.

4. RESULTS AND DISCUSSION

This study examined the impact of object detection methods on sign language recognition accuracy in complex backgrounds. Previous research has focused on feature extraction and classification but has not explored the combination of these with object detection in challenging environments. As shown in Table 3, both Faster RCNN and YOLOv5 effectively detect right-hand, left-hand, and face objects in greenscreen and complex backgrounds. YOLOv5 outperformed Faster RCNN in both settings, achieving higher mAP scores and faster processing times. Subsequent measurements focused on feature extraction.

Table 3. Object detection result

Dataset	mAP		Time	
	Faster RCNN	YOLOv5	Faster RCNN	YOLOv5
Greenscreen background	71.9%	74.1%	0.0236 s	0.0125 s
Complex background	71.3%	74.1%	0.0251 s	0.0138 s

During feature extraction with MobileNetV2, Faster RCNN showed decreased accuracy, while YOLOv5's accuracy improved. The baseline method, which skips object detection, performed best on greenscreen backgrounds but YOLOv5 excelled in complex backgrounds. The baseline model was also the fastest due to the absence of object detection. Detailed results are in Table 4.

Table 4. Feature extraction result

Experiments	Greenscreen background		Complex background	
	Accuracy	Time	Accuracy	Time
Skin segmentation + MobileNetV2	78.52%	0.0024s	69.14%	0.0028s
Faster RCNN + skin segmentation + MobileNetV2	67.14%	0.0256s	64.16%%	0.0271s
YOLOv5 + skin segmentation + MobileNetV2	75.41%	0.0145s	76.42%	0.0158s

In the classification stage using the LSTM model, the baseline method performs well on datasets with greenscreen backgrounds. However, its accuracy drops significantly when the method is implemented on complex backgrounds. Regardless, for the two methods implementing object detection, it is observed that the model classifies well on both types of datasets. This indicates that incorporating an object detection model at the beginning of the process successfully maintains good accuracy on both datasets.

Table 5 shows that both object detection models effectively addressed the challenges faced by the baseline model on complex datasets. SAcc, WER, BLEU, and SacreBLEU were calculated to evaluate performance at word and sentence levels. The baseline model performed poorly, while the Faster RCNN model showed improvement. The YOLOv5 model achieved the best results, with an SAcc of 49.29%, WER of 16.42%, BLEU of 40.97%, and SacreBLEU of 67.77%.

Table 5. Classification result

Experiments	Greenscreen background		Complex background	
	Accuracy	Time	Accuracy	Time
Skin segmentation + MobileNetV2 + LSTM	86.79%	0.0054s	44.88%	0.0058s
Faster RCNN + skin segmentation + MobileNetV2 + LSTM	77.71%	0.0286s	73.04%	0.0301s
YOLOv5 + skin segmentation + MobileNetV2 + LSTM	83.43%	0.0175s	83.58%	0.0188s

The study conducts a one-way ANOVA significance test to confirm the significance of the observed differences. It compares the results from each experiment. If the p-value is less than the alpha of 0.05, it indicates a significant difference between the two methods and vice versa.

Based on the findings from Table 6 and the significance test results in Table 7, it is clear that there is a significant increase in SAcc, BLEU, and SacreBLEU by the YOLOv5 model compared to the Faster RCNN and the baseline model. Simultaneously, there is a significant decrease in WER produced by the YOLOv5 model compared to the other two models. In terms of computational time, YOLOv5 shows a significant difference from the Faster RCNN model, which means YOLOv5 is significantly faster than the Faster RCNN model. This implies that YOLOv5 achieved the best results of the three models. Table 8 shows several of the results of the experiments.

Table 6. SAcc, WER, BLEU, and SacreBLEU evaluation on complex background

Experiments	SAcc	WER	BLEU	SacreBLEU
Skin segmentation + MobileNetV2 + LSTM	5.71%	54.07%	5.62%	17.37%
Faster RCNN + skin segmentation + MobileNetV2 + LSTM	35.00%	26.81%	23.76%	45.86%
YOLOv5 + skin segmentation + MobileNetV2 + LSTM	49.29%	16.42%	40.97%	67.77%

Table 7. One-way ANOVA significance test on complex background

Comparison	SAcc	WER	Time	BLEU	SacreBLEU
Baseline vs YOLOv5	Significant	Significant	Significant	Significant	Significant
Baseline vs Faster RCNN	Significant	Significant	Significant	Significant	Significant
Faster RCNN vs YOLOV5	Significant	Significant	Significant	Significant	Significant

Table 8. Samples of predicted Bisindo sign language gestures in a complex background

No	Actual label	Predicted label baseline model	Predicted label Faster RCNN model	Predicted label YOLOv5 model
1	26-125-97-78-85-100-47-132-105	26-93-26-78-24-24-47-125-36 (**WER: 0.67**)	26-125-125-78-85-125-47-132-105 (**WER: 0.22**)	26-125-147-78-85-100-47-132-105 (**WER: 0.11**)
2	103-109-120	130-109-120 (**WER: 0.33**)	70-109-120 (**WER: 0.33**)	103-109-120 (**WER: 0**)
3	70-114-18-39-8-9-12-4-7-56	56-114-18-39-8-9-18-4-25-18 (**WER: 0.4**)	70-114-18-39-8-9-54-4-54-56 (**WER: 0.2**)	70-114-18-39-8-9-12-4-7-56 (**WER: 0**)
4	70-144-121-128	70-70-41-128 (**WER: 0.5**)	70-144-121-125 (**WER: 0.25**)	70-144-121-128 (**WER: 0**)
5	125-60-148-94-132-105	60-70-105-59-94-105 (**WER: 0.83**)	125-113-148-94-132-105 (**WER: 0.17**)	125-60-148-103-132-105 (**WER: 0.17**)

In addition to the results above, there are several findings from the experiment. Out of 152 words in the dataset, 82 words were predicted correctly with 100% accuracy, and only one could not be predicted at all. The word is "*Eh*". The word "*Eh*" in Indonesian is a type of greeting word. This word could not be predicted because its appearance was only in one video, so the system did not sufficiently train it. Additionally, there are similar or synonymous words that can be predicted interchangeably, such as the word "*Cantik*", which means "Beauty," and the word "*Tampan*", which means "Handsome". Both words also have similar sign displays, so the system can incorrectly predict "*Cantik*" as "*Tampan*" or vice versa.

Another finding is related to the SAcc value. Based on the SAcc formula defined in the previous chapter, the SAcc value is greatly influenced by the length of the sentence. Simply put, no matter how long the sentence is, if there is even one word that is not predicted correctly, the SAcc value for that sentence is 0. This is why even the best model still has an SAcc value of only 49.29%.

SAcc and BLEU have a similar trend in evaluating this system. Both, when compared to SacreBLEU, show that SacreBLEU tends to achieve better results than SAcc and BLEU. This indicates the difference between the two metrics in evaluation. SAcc is not suitable for evaluating Bisindo because Bisindo does not have standard rules like SIBI. For example, Bisindo can have several variations of signs in terms of word order when expressing a sentence. If evaluated using SAcc, these variations will be considered incorrect, whereas they should be considered correct.

Analysis of the SAcc results found that SAcc is relatively good when the predicted sentence length is less than five words but relatively poor for sentences with a length of five words or more. For sentences with a length of less than five words, the average SAcc value is 62.67%, while for sentences with a length of five words or more, it is 34.85%. The SAcc value for sentences with a length of less than five words is similar to the SacreBLEU result, which is 67.77%. This further strengthens the argument that SacreBLEU is the appropriate metric for evaluating the Bisindo sign language recognition system, as its evaluation is more consistent.

5. CONCLUSION

This research aimed to solve the problem of recognizing Bisindo sign language in complex backgrounds and determine the suitable evaluation method. Experiments were conducted to address this issue. It was found that adding an object detection method at the beginning of the process could improve performance by first extracting hand and face objects. The best-performing object detection model was YOLOv5, achieving SAcc 49.29%, SacreBLEU 67.77%, BLEU 40.97%, and WER 16.42% on complex backgrounds, with a computation time of 0.0188 seconds per frame. SAcc was effective for sentences of less than five words, with an average SAcc of 62.67% for YOLOv5. However, for longer sentences, SAcc dropped to 34.85%, showing inconsistency. SacreBLEU, on the other hand, maintained consistency with a final score of 67.77% for all sentence lengths, making it the appropriate evaluation method for Bisindo sign language recognition.

APPENDIX

Table 2. Sentences in the dataset

No	Sentence in Indonesian	Sentence in English	Sentence in Bisindo	Sentence in Bisindo (translated into English)
1	*Besok saya mau pergi ke Bandung.*	Tomorrow, I am going to Bandung	*Mau – pergi – Bandung – besok*	Want – go – Bandung – tomorrow
2	*Kamu tidak boleh pulang sekarang.*	You are not allowed to go home now!	*Kamu – pulang – sekarang tidak boleh*	You – go home – now – do not
3	*Setiap pagi saya berjalan selama 30 menit.*	Every morning, I walk for 30 minutes.	*Saya – jalan-kaki – tiga-puluh – menit – pagi – pagi – pagi*	I – walk – thirty – minute – morning – morning – morning
4	*Karena takut, dia lari.*	Out of fear, He ran	*Dia – takut – dia – lari*	He – fear – he – run
5	*Ibu saya biasanya tidur jam 10 malam.*	My mother usually goes to sleep at 10 PM.	*Ibu – saya – tidur – jam -sepuluh – malam – biasanya*	My Mother – sleep – ten o'clock – night – usually
6	*Karena senang, dia tersenyum.*	She is smiling because She is happy.	*Dia senang dia senyum*	She – happy – She – smile
7	*Anak perempuan itu sedih karena ayahnya meninggal.*	The girl is sad because her father passed away.	*Bapak meninggal anak perempuan itu sedih*	Father – passed away – girl – sad
8	*Saya marah karena saya merasa dihina oleh dia.*	I am angry because I feel insulted by him/her.	*Saya merasa dia hina marah*	I – feel – He – insult – angry
9	*Jangan khawatir saya pasti membantu kamu.*	Don't worry, I will definitely help you.	*Khawatir jangan saya bantu pasti*	Worry - do not - i - will - help
10	*Saya bingung mengapa saya tidak boleh makan semangka.*	I'm confused why I'm not allowed to eat watermelon.	*Saya - makan - semangka tidak boleh - bingung - kenapa*	I - eat - watermelon - do not - confuse - why
11	*Setiap pagi ayah saya minum kopi dan makan nasi goreng.*	Every morning, my father drinks coffee and eats fried rice.	*Bapak saya minum kopi makan nasi goreng pagi pagi pagi*	My father - drink - coffee - eat - fried rice - morning - morning - morning

Table 2. Sentences in the dataset *(Continue…)*

No	Sentence in Indonesian	Sentence in English	Sentence in Bisindo	Sentence in Bisindo (translated into English)
12	*Adhi membeli buku itu di Hongkong.*	Adhi bought that book in Hong Kong.	*A D H I buku itu beli Hongkong*	A - D - H - I - that book - buy - Hongkong
13	*Saya ingin menjual sepeda motor saya.*	I want to sell my motorcycle.	*Motor saya ingin jual*	Motorcycle - i - want - sell
14	*Kakak saya memberi saya hadiah.*	My sister gave me a gift.	*Hadiah kakak beri*	Gift - sister - give
15	*Apakah kamu sudah menerima surat dari presiden Jokowi?*	Have you received a letter from President Jokowi?	*Surat Presiden Jokowi terima sudah*	Letter - President Jokowi - Received
16	*Penjahat itu memukul saya berkali-kali.*	The criminal hit me several times.	*Orang penjahat itu pukul-pukul saya*	Criminal - hit - me
17	*Saya tidak pernah mengerti pertanyaan anda.*	I never understood your question.	*Kamu pertanyaan tidak-mengerti*	You - question - never understood
18	*Berita itu sudah dilihat oleh ribuan orang.*	That news has been seen by thousands of people.	*Berita orang ribuan nonton sudah*	News - people - thousands - seen
19	*Apakah kamu pernah membaca nover bahasa inggris?*	Have you ever read an English novel?	*Buku Bahasa Inggris baca sudah pernah*	Book - english - read
20	*HP ini dibeli oleh Laura dengan harga tiga juta.*	This mobile phone was bought by Laura for three million.	*Handphone ini L A U R A beli harga tiga juta*	Mobile phone - this - L - A - U - R - A - buy - price - three - million
21	*Pencuri itu dipukul oleh polisi.*	The thief was beaten by the police.	*Orang pencuri itu polisi pukul*	Thief - police - beaten
22	*Kambing itu dimakan ular*	The goat was eaten by a snake.	*Kambing ini makan ular*	Goat - eaten - snake
23	*Obat ini harus diminum tiga kali sehari*	This medicine should be taken three times a day.	*Minum-obat harus satu hari tiga kali*	Take medicine - should - three times - a day
24	*Dia bertanya kepada ku di mana saya lahir*	He asked me where I was born.	*Dia tanya saya lahir mana*	He - ask - me - born - where
25	*Apa yang sedang kamu pikirkan?*	What are you thinking about?	*Kamu pikir apa*	You - think - what
26	*Di mana anda tinggal?*	Where do you live?	*Kamu tinggal mana*	You - live - where
27	*Kemana kamu mau pergi?*	Where do you want to go?	*Kamu mau pergi mana*	You - want to - go - where
28	*Mengapa kemarin kamu tidak kuliah?*	Why didn't you attend the class yesterday?	*Kemarin kamu kuliah tidak kenapa*	Yesterday - you - class - did not - attend
29	*Bagaimana keadaan ibumu?*	How is your mother?	*Ibu kamu bagaimana sehat*	Mother - you - how - well
30	*Jam berapa kita istirahat?*	What time do we take a break?	*Kita istirahat jam berapa*	We - take a break - what time
31	*Kapan kamu belajar Bisindo?*	When did you learn Bisindo?	*Kamu belajar Bisindo kapan*	You - learn - Bisindo - when
32	*Badan ku gemuk, tapi badan adik ku kurus*	My body is fat, but my brother's body is thin.	*Badan saya gemuk tapi badan adik kurus*	My body - fat - but - my brother's body - thin
33	*Masakan padang itu enak, tetapi mahal*	Padang cuisine is delicious, but expensive.	*Makan padang itu enak tapi mahal*	Padang cuisine - delicious - but - expensive
34	*Dia anak baik sehingga banyak orang menyukainya*	He is a good kid, so many people like him.	*Dia anak baik orang orang suka suka*	He - kid - good - many people - like - him
35	*Artis korea itu sangat tampan*	The Korean celebrity is very handsome.	*Artis korea itu tampan*	Celebrity - korean - handsome
36	*Penyanyi cantik itu bisa bermain gitar*	The beautiful singer can play the guitar.	*Orang penyanyi cantik itu bisa bermain-gitar*	Singer - beautiful - can - play guitar
37	*Tulisannya sangat jelek sehingga tidak bisa saya baca*	His handwriting is very bad, so I can't read it.	*Dia tulisan baca tidak-bisa*	He - handwriting - read - can not
38	*Iwan memakai baju merah, sedangkan Adhi memakai baju putih*	Iwan is wearing a red shirt, while Adhi is wearing a white shirt.	*I W A N baju pakai-baju merah A D H I baju pakai-baju putih*	I - W - A - N shirt - wear - red - A - D - H - I - shirt - wear - white
39	*Tolong, matikan AC-nya!*	Please, turn off the AC!	*A C mati tolong*	A - C - turn off - please
40	*Jangan duduk di atas meja!*	Don't sit on the table!	*Meja duduk jangan*	Table - sit - do not

ACKNOWLEDGEMENTS

This work is supported by the Tokopedia-UI AI Center of Excellence and the Funding Program for Innovation at the University of Indonesia 2023 Program P3 No: 33/PKS/WR III-DISTP/UI/2023 No: 0026/SPIC/ES/2023.

REFERENCES

[1] R. Rastgoo, K. Kiani, and S. Escalera, “Sign language recognition: a deep survey,” *Expert Systems with Applications*, vol. 164, p. 113794, Feb. 2021, doi: 10.1016/j.eswa.2020.113794.

[2] L. Zheng, B. Liang, and A. Jiang, “Recent advances of deep learning for sign language recognition,” in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov. 2017, pp. 1–7, doi: 10.1109/DICTA.2017.8227483.

[3] K. Nimisha and A. Jacob, “A brief review of the recent trends in sign language recognition,” in *2020 International Conference on Communication and Signal Processing (ICCSP)*, Jul. 2020, pp. 186–190, doi: 10.1109/ICCSP48568.2020.9182351.

[4] H. M. Mariappan and V. Gomathi, “Real-time recognition of Indian sign language,” in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Feb. 2019, pp. 1–6, doi: 10.1109/ICCIDS.2019.8862125.

[5] Q. Xiao, M. Qin, P. Guo, and Y. Zhao, “Multimodal fusion based on LSTM and a couple conditional hidden Markov model for Chinese sign language recognition,” *IEEE Access*, vol. 7, pp. 112258–112268, 2019, doi: 10.1109/ACCESS.2019.2925654.

[6] S. M. S. Shah, H. Abbas Naqvi, J. I. Khan, M. Ramzan, Zulqarnain, and H. U. Khan, “Shape based Pakistan sign language categorization using statistical features and support vector machines,” *IEEE Access*, vol. 6, pp. 59242–59252, 2018, doi: 10.1109/ACCESS.2018.2872670.

[7] R. D. RAJ and A. JASUJA, “British sign language recognition using HOG,” in *2018 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS)*, Feb. 2018, pp. 1–4, doi: 10.1109/SCEECS.2018.8546967.

[8] R. Rastgoo, K. Kiani, and S. Escalera, “Hand sign language recognition using multi-view hand skeleton,” *Expert Systems with Applications*, vol. 150, p. 113336, Jul. 2020, doi: 10.1016/j.eswa.2020.113336.

[9] S. Baek, K. I. Kim, and T.-K. Kim, “Augmented skeleton space transfer for depth-based hand pose estimation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8330–8339, doi: 10.1109/CVPR.2018.00869.

[10] A. Spurr, J. Song, S. Park, and O. Hilliges, “Cross-modal deep variational hand pose estimation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 89–98, doi: 10.1109/CVPR.2018.00017.

[11] S. He, “Research of a sign language translation system based on deep learning,” in *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, Oct. 2019, pp. 392–396, doi: 10.1109/AIAM48774.2019.00083.

[12] O. B. Hoque, M. I. Jubair, M. S. Islam, A.-F. Akash, and A. S. Paulson, “Real time Bangladeshi sign language detection using faster R-CNN,” in *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, Dec. 2018, pp. 1–6, doi: 10.1109/CIET.2018.8660780.

[13] S. Kim, Y. Ji, and K.-B. Lee, “An effective sign language learning with object detection based ROI segmentation,” in *2018 Second IEEE International Conference on Robotic Computing (IRC)*, Jan. 2018, pp. 330–333, doi: 10.1109/IRC.2018.00069.

[14] H. D. Alon, M. A. D. Ligayo, M. P. Melegrito, C. Franco Cunanan, and E. E. Uy II, “Deep-hand: a deep inference vision approach of recognizing a hand sign language using American alphabet,” in *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Mar. 2021, pp. 373–377, doi: 10.1109/ICCIKE51210.2021.9410803.

[15] S. Daniels, N. Suciati, and C. Fathichah, “Indonesian sign language recognition using YOLO method,” *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, p. 012029, Feb. 2021, doi: 10.1088/1757-899X/1077/1/012029.

[16] M. Faisal, “Development of Indonesian sign language (Bisindo) movement recognizer model using MobileNetV2 as a feature extractor and LSTM as a classifier,” Universitas Indonesia, 2021.

[17] N. F. P. Setyono and E. Rakun, “Recognizing word gesture in sign system for Indonesian language (SIBI) sentences using DeepCNN and BiLSTM,” in *2019 International Conference on Advanced Computer Science and information Systems (ICACSIS)*, Oct. 2019, pp. 199–204, doi: 10.1109/ICACSIS47736.2019.8979772.

[18] M. H. Nugraha and E. Rakun, “Solving complex background problem using retinanet for sign system for Indonesian language (SIBI) gesture-to-text translator,” in *2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2022, pp. 45–52, doi: 10.1109/ICACSIS56558.2022.9923450.

[19] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[20] T. F. Dima and M. E. Ahmed, “Using YOLOv5 algorithm to detect and recognize American sign language,” in *2021 International Conference on Information Technology (ICIT)*, Jul. 2021, pp. 603–607, doi: 10.1109/ICIT52682.2021.9491672.

[21] M. Post, “A call for clarity in reporting BLEU scores,” in *WMT 2018 - 3rd Conference on Machine Translation, Proceedings of the Conference*, 2018, vol. 1, pp. 186–191, doi: 10.18653/v1/w18-6319.

[22] N. Padmaja, B. N. S. Raja, and B. P. Kumar, “Real time sign language detection system using deep learning techniques,” *Journal of Pharmaceutical Negative Results*, vol. 13, no. S01, pp. 1052–1059, Jan. 2022, doi: 10.47750/pnr.2022.13.s01.126.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.

[25] R. F. Rahmat, T. Chairunnisa, D. Gunawan, and O. S. Sitompul, “Skin color segmentation using multi-color space threshold,” in *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, Aug. 2016, pp. 391–396, doi: 10.1109/ICCOINS.2016.7783247.

[26] R. Padilla, S. L. Netto, and E. A. B. da Silva, “A survey on performance metrics for object-detection algorithms,” in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Jul. 2020, pp. 237–242, doi: 10.1109/IWSSIP48289.2020.9145130.

[27] A. C. Morris, V. Maier, and P. Green, “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition,” in *Interspeech 2004*, Oct. 2004, pp. 2765–2768, doi: 10.21437/Interspeech.2004-668.

BIOGRAPHIES OF AUTHORS

Muhammad Alfhi Saputra received his bachelor of degree in Faculty of Informatics from the Telkom University, Bandung, Indonesia in 2021. From 2021 until now he is a master's student in the Faculty of Computer Science in University of Indonesia. His research interests include, artificial intelligence, deep learning, and computer vision. He can be contacted at email: muhammad.alfhi@ui.ac.id.

Erdefi Rakun received her bachelor degree in Electrical Engineering from the University of Indonesia, in Jakarta, Indonesia in 1982. She received her M. Sc. in Computer Science from University of Minnesota, USA, 1988. She received her Ph.D. in Computer Science from University of Indonesia, in 2017. From 1986 until now, she is a full-time lecturer in the faculty of Computer Science in University of Indonesia, holding an Academic rank of Associate Professor. Her research interests include, machine learning, deep learning, image processing for Indonesian sign language recognition systems. She can be contacted at email: erdefi.rakun@cs.ui.ac.id.