JurnalIlmuKomputer dan Informasi (Journal of Computer Science and Information)
15/2 (2022), 69-79. DOI: http://dx.doi.org/10.21609/jiki.v15i2.1014

Improving Recognition of SIBI Gesture by Combining Skeleton and Hand Shape Features

Erdefi Rakun[1] and Noer Fitria Putra Setyono[2]

[1, 2]Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia

E-mail: [1]erdefi.rakun@cs.ui.ac.id, noer.fitria@ui.ac.id

Abstract

SIBI (Sign System for Indonesian Language) is an official sign language system used in school for hearing impairment students in Indonesia. This work uses the skeleton and hand shape features to classify SIBI gestures. In order to improve the performance of the gesture classification system, we tried to fuse the features in several different ways. The accuracy results achieved by the feature fusion methods are, in descending order of accuracy: 88.016%, when using sequence-feature-vector concatenation, 85.448% when using Conneau feature vector concatenation, 83.723% when using feature-vector concatenation, and 49.618% when using simple feature concatenation. The sequence-feature-vector concatenation techniques yield noticeably better results than those achieved using single features (82.849% with skeleton feature only, 55.530% for the hand shape feature only). The experiment results show that the combined features of the whole gesture sequence can better distinguish one gesture from another in SIBI than the combined features of each gesture frame. In addition to finding the best feature combination technique, this study also found the most suitable Recurrent Neural Network (RNN) model for recognizing SIBI. The models tested are 1-layer, 2-layer LSTM, and GRU. The experimental results show that the 2-layer bidirectional LSTM has the best performance.

Keywords: *SIBI, Long Short Term Memory, Gated Recurrent Unit, Feature Concatenation*

1. Introduction

Everyday life is harder for those with the hearing impaired. Not only do they have difficulty in communicating with those without hearing impairment, but they also need a different system to communicate with those with hearing loss. Vision-based communication techniques in sign language and lip-reading can mitigate these communication difficulties. Sign language consists of the finger, hand, movement, body, and facial movements that, when combined, represent the word the signer wants to convey [1].

SIBI is a sign language system that conforms to the grammatical structure of Indonesian, including its usage of inflection words [2]. Affixes are an essential part of Indonesian since many ideas are represented by imbuing a root word with more affixes (prefixes, suffixes, circumfixes, and infixes). This treatment applies to nouns, verbs, and adjectives, and as such, any sign language system that hopes to emulate this trait of Indonesian uses a similar construct in its gestures. The addition of this affix serves to give additional meaning to the root word. For example, the prefix "me" + root word "lempar" (=to throw) + the suffix "i" gives the meaning of "throwing at " to a stationary object. Furthermore, the prefix "me" + root word "lempar" (=to throw) + the suffix "kan" give the meaning of "throwing with" with an active object.

Unfortunately, not many people are proficient in SIBI. As a result, the deaf still has difficulties in communicating. This study aims to create an application to bridge this communication barrier by translating SIBI gestures into text. The application must be able to distinguish each SIBI gesture quickly and precisely. Gestures videos are significant; therefore, processing them will require heavy computation. For the gesture recognition process to be carried out quickly and precisely, we need a technique to get only some information but enough to distinguish each gesture.

Generally, every gesture has unique hand shapes and arm movements [1]. This uniqueness can be used as a distinguishing feature between one gesture from another. This study uses two features extraction techniques: the hand shape feature, which stores hand shape information, and the skeleton feature that stores arm movement information. The skeleton and the handshape features complement each other. The usage of the skeleton features alone results in the misclassification of gestures with the same hand movements but different hand shapes. This misclassification of gestures happens quite often in prefix and suffix gestures. Conversely, if only the hand shape features are used, gestures with the same handshapes but different hand movements are misclassified.

It is necessary to see examine how prefix gestures are formed in SIBI. The right-hand fingers form the first letter of the prefix according to the provisions of the SIBI alphabet. For example, at the prefix "me," the right-hand finger will form the letter "m." Furthermore, the left-hand palm is upright facing to the right at each prefix. The two palms will meet in front of the chest. Fig. 1 shows the right and left-hand gestures for the "me," "se," and "te" prefixes in SIBI.

Furthermore, the suffix gesture depends only on the right hand. The left-hand stands still by the side of the body. Like the prefix, the right-hand finger also forms the first letter of the suffix. For example, at the suffix "lah," the right-hand finger will show the letter "l"; at the suffix "kan," the right-hand fingers form the letter "k." Fig. 2 shows the "kan" and "lah" gestures.

Several SIBI gestures have similar handshapes, but the position and direction of hand movement are different. For example, the gestures of the word "bibi" (= aunty) and the word "biru" (= blue") have the same right-hand shape. Fig. 3 shows the gestures of "bibi" and "biru." The right-hand position of the word "bibi" is next to the right ear, while the word "blue," is in front of the right shoulder. The difference in the right-hand position causes these two words to have different skeleton features [1].

The following studies demonstrate the need to use both the skeleton and handshape features in a sign language recognition system. Research by [3] that used a skeleton could only recognize gestures with the right hand only. [4] who only uses hand shape data, can only recognize the alphabet. [5] using Image-based Hand gesture recognition can only recognize numeric gestures from 1 to 5. These three studies show that using one feature 0only will limit the gesture that can be recognized. Our previous research [6],[7] proved that concatenate skeleton features and hand shapes can recognize isolated gestures with an accuracy upto 95.4%. The weakness of [7] is that the method proposed is not good at recognizing continuous gestures. [8] research also applies the combination of hand and skeleton features obtained from the Kinect camera, leap motion, and the handy cam. This research can recognize the dynamic and static signs for independent signer mode with 88.09% accuracy. The feature extraction technique offered by [8] cannot be used in our study due to differences in data formats. The data that we have is a 2D image taken by a mobile phone camera.

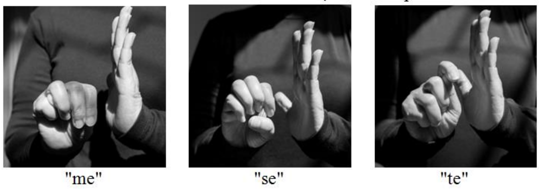


Fig. 1. Similar skeleton movement with different hand shape among prefixes [1]

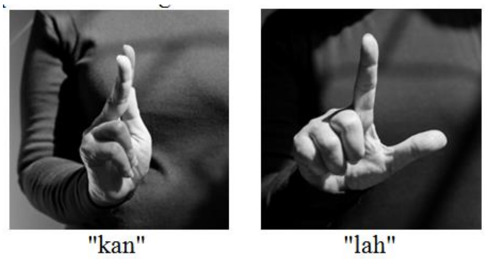


Fig. 2. Similar skeleton movement with different hand shape among suffixes [1]

This study aims to find a way to combine these two features in order to facilitate accurate SIBI gesture to Indonesian text translation. The techniques commonly used to combine several types of features are the concatenation technique or using multi-channel. There are many ways to concatenate the features, such as a simple one by combining all features into one group [6], [9]; or a complex one by combining preprocessed features [8], [9],[10],[12]. In multi-channel, each feature will be processed on a separate channel or track. Each channel has its own feature extraction technique and a model for recognizing input in that channel. Then there will be a process that combines the outputs of each track and makes decisions based on the combined outputs [13], [14].

There are three considerations to combine the skeleton and hand shape features:

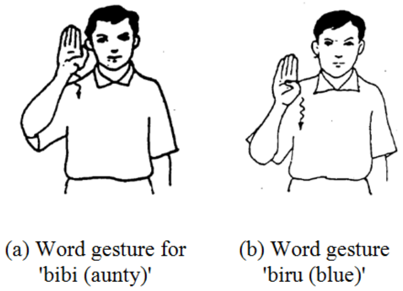

(a) Word gesture for 'bibi (aunty)'
(b) Word gesture 'biru (blue)'

Fig. 3. Similar hand shape with different skeleton movement among words [1]

1. We are building a mobile application that can translate SIBI gestures into Indonesian language text, and feature combination is one component in the mobile application. Features combining process must be feasible to be carried out on a mobile phone processor, which is relatively less powerful than a computer.
2. A mobile application should be able to translate gestures to text in real-time. For this reason, the application needs a fast feature combination technique.
3. Our skeleton feature consists of only 4 variables, while the hand shape consists of 112 variables. Therefore, we need a feature combination technique that can overcome this imbalance length problem.

Based on these considerations, we examine four feature concatenation techniques. We did not try the multi-channel approach due to its high computational cost.

The remainder of this paper is organized as follows. Section 2 states the experimental flow, dataset, and feature extraction techniques. Section 3 explains about Single Feature Prediction results as the baseline case. Section 4 discusses four Features concatenation techniques. Section 5 analyzes the experiment results. Finally, section 6 closes this paper with the conclusion and future works.

2. Methodology

This section will discuss the experimental flow, the dataset used, how to extract the skeleton and hand shape features, and the computer used in this study.

2.1. Experiment flow

The experiment flow is shown in Fig. 4 below. The first process performs feature extraction of each frame from the input video. The video input is a recording of the SIBI gesture taken by the camera of the Samsung S9+ smart phone. The features generated by this first process are stored in 2 files, one for the skeleton and another one for the hand shape feature. The second process is to get the baseline case. This process measures the accuracy of SIBI gesture recognition when using a single feature, the skeleton feature or the hand shape feature only. In this process, we also tried to find the best RNN model for the dataset used. In the next stage (processes 3-6), we look for feature concatenation techniques to improve the system's ability to recognize SIBI signals. The last process evaluates all experimental results, both in terms of accuracy and processing time.

2.2. Dataset

The dataset is considered primary data, as our group obtained it from the ground up. Three teachers and two deaf students from SLB Santi Rama, the special-needs school, performed the gestures. For the footage to represent real-life use cases, the dataset is recorded by the onboard camera of a Samsung S9+.

The sentences to be performed were chosen based on the input from the Santi Rama teachers.

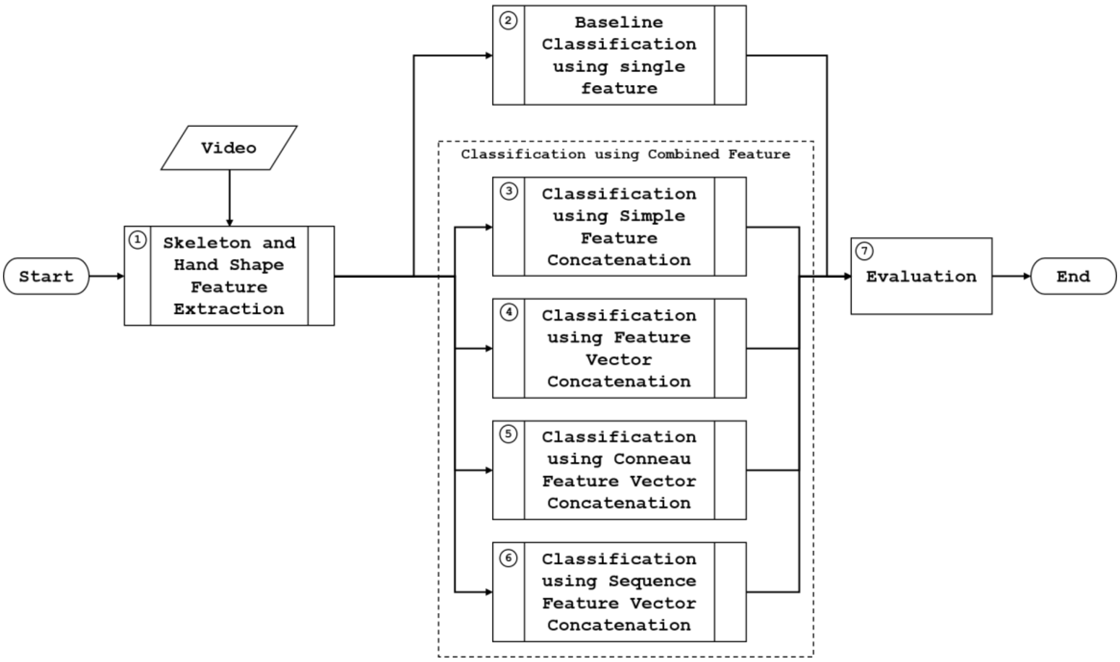

Start
Video
① Skeleton and Hand Shape Feature Extraction
② Baseline Classification using single feature
Classification using Combined Feature
③ Classification using Simple Feature Concatenation
④ Classification using Feature Vector Concatenation
⑤ Classification using Conneau Feature Vector Concatenation
⑥ Classification using Sequence Feature Vector Concatenation
⑦ Evaluation
End

Fig. 4. Experiment Flow

These sentences form a large cross-section of what the deaf need when they are in public places (such as banks, hospitals, theatres, department stores), taking public transport, as well as the typical greetings and introductions. Table 1 shows all the sentences in the dataset.

The Gesture Recognition system that we built will recognize the words in the sentence one by one according to the order in which they appear in the sentence. The classifier RNN needs to be trained to recognize each word in the dataset. We tried to represent the occurrence of each word in the dataset equally by recording each sentence 5-25 times.

Table 2 shows the number of samples for each word type in the dataset. The first column in Table 2 contains the word groups in SIBI. The second column contains the number of word labels in the dataset in each group, and the third column is the number of data per group. The data is divided into five-folds, four are used in training, and one is reserved as the validation set. The same folds are used in all the experiments.

Table 2. Type of gestures in the dataset

Group of Word	Total Words/Group	Number of Data/group
Root Word	69	10775
Prefix	3	942
Suffix	5	1214
Question Word	4	1013
Preposition	2	832
Total	83	14776

2.3. Overview of the Skeleton and Handshape Feature

2.3.1 Skeleton Feature

The extraction of the skeleton features involves representing the speaker in each (image) frame into four angles. θ_1 represents the angle of the upper arm relative to the shoulder, whereas θ_2 represents the angle of the forearm relative to the elbow. θ_3, θ_4 represent the same angles for the left hand, respectively (Fig. 5).

The data used in [9] consists solely of root words that are different from the SIBI dataset used in this work, containing sentences. There are several improvements to the skeleton features relative to the work done in [9]:

- Skin detection now uses HSV to YCrCb, along with a tweak in the thresholds.
- Facial detection uses only the upper half of the frame.
- The green threshold is now changed.

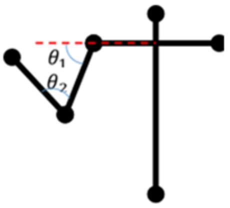
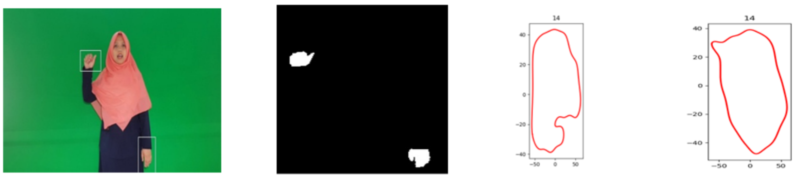


Fig. 5. Skeleton Feature

Table 1. List of sentences in the dataset

No	Sentence	Repetition
1	Siapa namamu? (What's your name?)	50
2	Di mana alamat rumahmu? (Where do you live?)	75
3	Di mana sekolahmu? (Where is your school located?)	75
4	Bolehkah saya minta nomor teleponmu? (May I have your telephone number?)	75
5	Film apa yang sedang diputar? (What movies are playing right now?)	75
6	Jam berapa film ini diputar? (At what times will this movie be shown?)	50
7	Berapa harga karcis film ini? (How much would a ticket for this movie cost?)	75
8	Di mana film ini diputar? (Where is this movie being shown?)	75
9	Apa nama sayuran itu? (What is that vegetable called?)	25
10	Berapa harga sayuran itu? (How much does that vegetable cost?)	50
11	Apakah harga sayuran ini boleh ditawar? (Is the price of this vegetable negotiable?)	75
12	Berapa jumlah yang harus saya bayar? (So how much do I have to pay for all of this?)	75
13	Kami ingin pergi ke kota tua, naik bis apa? (We would like to go to the Old Town, which bus do we have to take?)	100
14	Berapa harga karcis yang harus saya bayar? (How much would the tickets cost?)	75
15	Kami harus turun di mana? (At which station should we get off?)	75
16	Adakah cara lain kita pergi ke kota tua? (Is there another way we can get to the Old Town?)	100
17	Saya ingin membuka tabungan, bagaimana caranya? (How would I go about opening a savings account?)	75
18	Bagaimana cara menabung? (How would I go about saving money?)	50
19	Di mana kami bisa mengambil tabungan? (Where can we withdraw our savings?)	75
20	Bagaimana cara mengirim uang melalui bank? (How can I send money through a bank?)	125
21	Selamat natal dan tahun baru (Merry Christmas and Happy New Year)	125
22	Selamat idul fitri mohon maaf lahir dan batin (Happy Eid ul-Fitr, please forgive me for my mistakes)	125
23	Selamat ulang tahun (Happy Birthday)	125
24	Semoga panjang umur (May you live for as long as you wish)	125
25	Saya sering sakit kepala, saya harus periksa ke bagian mana? (I frequently get headaches, which medical specialty department should I visit?)	100
26	Saya ingin ke dokter umum, siapa nama dokternya? (I want to see a general practitioner; What is the doctor's name?	75
27	Jam berapa dokter datang? (At what time is the doctor expected to arrive?)	25
28	Di apotek mana obat ini bisa dibeli? (At which pharmacy can I buy this medicine?)	75



(a)
Original frame

(b)
Segmentation Result

(c)
EFD for right hand

(d)
EFD for left hand

Fig. 6. Hand Shape feature extraction

We resolved the issue whereby a hand's centroid disappears when the hand goes off-frame using that hand's last known centroid position instead.

2.3.2. *Hand Shape Feature*

The extraction of the hand shape feature consists of determining three components: the segmentation of the hand blob area, the calculation of the hand blob centroid, and the coordinate of the hand's contour by using Elliptical Fourier Descriptor (EFD) of order 14 [9]. Each frame's hand shape features consist of 112 dimensions. An example of the extracted features can be seen in Fig. 6.

There are several improvements to the hand shape feature extraction process relative to the work done in [9]:

1. YCrCb Skin detection is now performed by examining the YCrCb components.
2. Face detection is now only performed on the upper half of the body in the frame.
3. An elliptical mask is now used to mask the face from the frame. Previous works used a rectangular mask, which results in the gesturing hand being masked on some occasions.
4. An additional routine is added to cover the scenario whereby the hand is inside the face's mask. This routine imputes the missing data in that frame by using the hand shape that immediately precedes the problematic frame.

3. Single Feature Prediction as Baseline case

3.1. Experiments using Skeleton and Handshape features in the dataset

There are 3 RNN models tested here: 1- and 2-layer LSTM, GRU, BiLSTM, and BiGRU. Table 3 shows the prediction accuracies of each experiment. The training and testing processes were run on a machine with an i7-7700 CPU, 32 GB of RAM, and a GTX 1060 GPU running Ubuntu 16.04.4 LTS.

3.2. Analysis

The LSTM and GRU models using each feature set alone performed very differently in the dataset. The best result for the skeleton feature alone is 82.849% accuracy using the 2-layer BiLSTM, whereas the best result utilizing the hand shape feature is 55.530% accuracy also by using the 2-layer BiLSTM. The spread of the accuracy for the different model architectures is very different for the two feature sets, with the spread for the skeleton feature set's accuracies being smaller than the hand shape feature. This fact means that the skeleton feature is the more stable and more decisive (more correlated to the labels given the model architecture) feature set. Moreover, the 2-layer

Table 3. RNN Prediction Accuracies by using Skeleton or Hand Shape Feature

Feature	**Layer**	**RNN Model**	**Fold**					
			1	**2**	**3**	**4**	**5**	**Average**
Skeleton	**1-Layer**	LSTM	75.678%	76.750%	**78.696%**	76.951%	76.192%	76.854%
		GRU	**80.235%**	79.129%	79.133%	77.688%	78.319%	**78.901%**
		BiLSTM	72.350%	73.869%	**76.882%**	72.730%	72.933%	73.753%
		BiGRU	**81.475%**	75.109%	78.730%	74.137%	77.599%	77.410%
	2-Layer	LSTM	68.679%	69.816%	70.430%	**71.357%**	69.708%	69.998%
		GRU	**68.130%**	67.638%	66.263%	65.896%	67.101%	67.006%
		BiLSTM	83.293%	**83.719%**	83.535%	81.709%	81.990%	**82.849%**
		BiGRU	83.225%	82.312%	**83.804%**	81.776%	81.818%	82.587%
Hand Shape	**1-Layer**	LSTM	8.748%	**9.983%**	8.132%	8.677%	6.827%	8.473%
		GRU	9.571%	8.375%	9.409%	10.620%	**12.281%**	10.051%
		BiLSTM	10.669%	9.715%	10.349%	**19.966%**	9.949%	12.130%
		BiGRU	19.348%	14.070%	18.011%	**26.700%**	20.480%	**19.722%**
	2-Layer	LSTM	28.885%	23.652%	25.706%	25.427%	**31.458%**	27.025%
		GRU	**45.489%**	26.566%	34.375%	32.060%	42.401%	36.178%
		BiLSTM	**59.828%**	51.323%	49.698%	57.487%	59.314%	**55.530%**
		BiGRU	44.151%	49.749%	**53.360%**	35.913%	50.292%	46.693%

bidirectional models exhibit clear superiority over the other model architectures in this dataset, regardless of which feature set is used.

The excellent performance of bidirectional RNN is likely due to the difference in the location of the differentiating gesture components for SIBI's root words, prefixes, and suffixes [15]. RNN will work well if it can immediately find the characteristics of the data/features that need to be analyzed. The SIBI's suffixes have their specific elements in the initial frames that match the forward RNN. On the other hand, the differentiating component of the root words and the prefixes are located in the final frames that fit the backward RNN (note: the root words include all words in SIBI that do not belong to prefixes and suffixes). The location of the differentiating gesture components for SIBI's gestures is why the Bidirectional RNN can recognize almost all types of hand movement in SIBI.

4. Features Concatenation Techniques

This section presents four concatenation techniques to improve single feature ability to recognize SIBI gestures: Simple, Feature Vector, Conneau, and Sequence-Feature-Vector Concatenation. Each technique will be discussed how it works, the accuracy obtained, and the analysis of the experimental results.

4.1. Simple Feature Concatenation (SFC) Model

This model concatenates four skeleton feature data and 112 hand shape feature data into a combined feature with a length of 116. RNN then uses this combined feature to recognize SIBI signals. Fig. 7 shows how this model works.

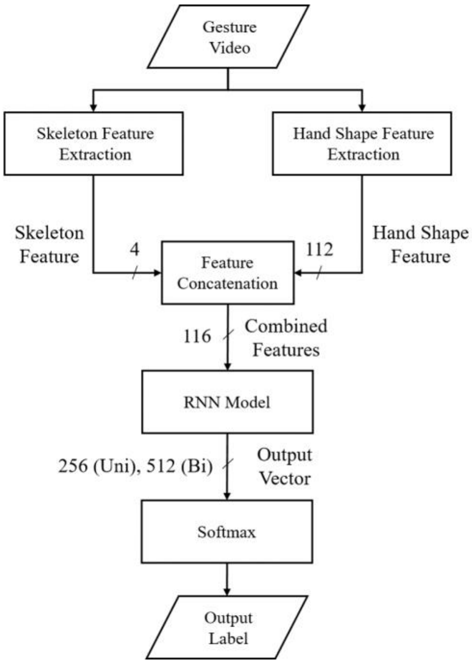


Fig. 7. Simple Feature Concatenation Model

The RNN models studied were 1-layer, 2-layer LSTM, GRU, Bidirectional LSTM, and Bidirectional GRU. Table 4 shows the accuracy obtained from each RNN model.

The dimensions of the features extracted are not evenly matched: the hand shape feature is 112-dimensional (per frame), and the skeleton feature is only 4-dimensional. Looking at recognition accuracies obtained by the RNN models in Tables 3 and 4 closely, it is shown that the combined feature recognition results are lower than a single feature, as the maximum average accuracy obtained by this as-is-concatenation model is 49.618%. This accuracy is less than the maximum average accuracy obtained using the skeleton features only (82.849%) and over the maximum average accuracy obtained using the hand shape features only (55.530%). The best performing model tested with this setup is the 2-layer BiLSTM.

From the results obtained above, it is clear that something has to be done about the relative weighting of the concatenated tensor. Concatenating the feature tensor as-is weighs the hand shape features disproportionately, as shown by the fact that the results are worse than when the skeleton features are used alone

4.2. Feature Vector Concatenation

In order to weigh the two features equally, it is decided to ensure that the concatenated feature tensor be composed of equal parts hand shape and equal parts skeleton features. Equal part means that there has to be some dimensionality expansion performed on the skeleton features.

Table 4. RNN Prediction Accuracies by using Simple Feature Concatenation Model

Feature	Layer	RNN Model	Fold					
			1	2	3	4	5	Average
Simple Feature Concatenation Model	**1-Layer**	LSTM	4.700%	**7.169%**	3.159%	5.293%	4.803%	5.025%
		GRU	4.974%	5.896%	4.906%	3.149%	**5.901%**	4.965%
		BiLSTM	9.434%	**10.184%**	9.711%	9.749%	8.954%	9.606%
		BiGRU	11.458%	17.856%	12.466%	8.911%	**20.892%**	14.317%
	2-Layer	LSTM	26.484%	23.283%	33.468%	**35.310%**	27.033%	29.115%
		GRU	**44.837%**	29.816%	44.422%	34.070%	27.376%	36.104%
		BiLSTM	41.269%	41.269%	49.832%	55.310%	**60.412%**	**49.618%**
		BiGRU	**53.962%**	46.231%	19.691%	52.261%	51.081%	44.645%

To take advantage of the fact that the feature vector is in the form of a sequence, we decided to use another pair of RNNs to perform this dimensionality expansion task. In the end, to keep the same level of complexity as in the previous works [16], and since the size of the hidden layer of the dimensionality-expander RNN determines the final dimension that it outputs, the size of the hidden layer was set to 256 in this case. It means that both features undergo dimensionality expansion as well as the sequence processing via back-propagation through time that is inherent in RNNs. The size of each expanded feature is *1 x 256* for uni-directional RNN (LSTM or GRU) and *1 x 512* for bi-directional RNN. The size of combined features is *1 x 512* for uni-directional RNN and *1 x 1024* for bi-directional RNN. The combined features are then fed straight into a softmax layer to determine which word the features correspond to. The model architecture can be seen in Fig. 8, while the experiment results are listed in Table 5.

The results show that the combined model is more accurate (83.723%) than the model trained on single feature sets alone (82.849% and 55.53% for the skeleton and hand shape features, respectively). The Bi-LSTM performed slightly better than the Bi-GRU.

4.3. Conneau Feature Vector Concatenation

In order to improve the performance even further, we tried to implement the architecture first proposed by [5], as shown in Fig. 9. Conneau's architecture is similar to the architecture in Fig. 8, apart from the concatenation method. Conneau proposed using a combination of three methods of feature tensor combination: element-wise product, element-wise difference, and concatenation. All three are used to create the combined feature tensor, using the outputs of RNN Model I (skeleton) and RNN Model II (hand shape) as constituents.

The rest of the tested configuration is the same, using a 2-layer Bi-LSTM and Bi-GRU. The results can be seen in Table 6. The best result is an accuracy of 85.448%, again higher than the models trained on the constituents alone. The 2-layer Bi-LSTM performed slightly better than 2-layer Bi-GRU.

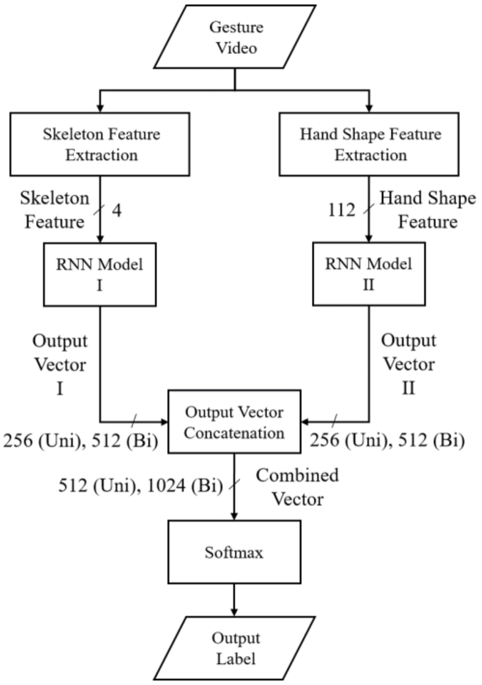


Fig. 8. Feature Vector Concatenation Model

The rest of the tested configuration is the same, using a 2-layer Bi-LSTM and Bi-GRU. The results can be seen in Table 6. The best result is an accuracy of 85.448%, again higher than the models trained on the constituents alone. The 2-layer Bi-LSTM performed slightly better than 2-layer Bi-GRU.

4.4. Proposed Method - The Sequence-Feature-Vector Concatenation Model

This proposed model architecture blends the same frame's skeleton and hand shape features. Combining these two features, which are different in predictive power and dimensionality, is performed over a sequence spanning 19 frames instead of frame by frame in the previous works. The operational flow of this model can be divided

Table 5. RNN Prediction Accuracies by using Feature Vector Concatenation

Feature	RNN Model	Fold					
		1	2	3	4	5	Average
Feature Vector Concatenation	2-Layer BiLSTM	83.911%	83.384%	**84.745%**	83.283%	83.293%	**83.723%**
	2-Layer BiGRU	82.710%	83.786%	**83.804%**	81.675%	83.259%	83.047%

Table 6. RNN Prediction Accuracies by using Conneau Feature Vector Concatenation

Feature	RNN Model	Fold					
		1	2	3	4	5	Average
Conneau Feature Vector Concatenation	2-Layer BiLSTM	85.901%	85.126%	85.249%	**86.231%**	84.734%	**85.448%**
	2-Layer BiGRU	82.367%	82.044%	**84.745%**	83.451%	83.979%	83.317%

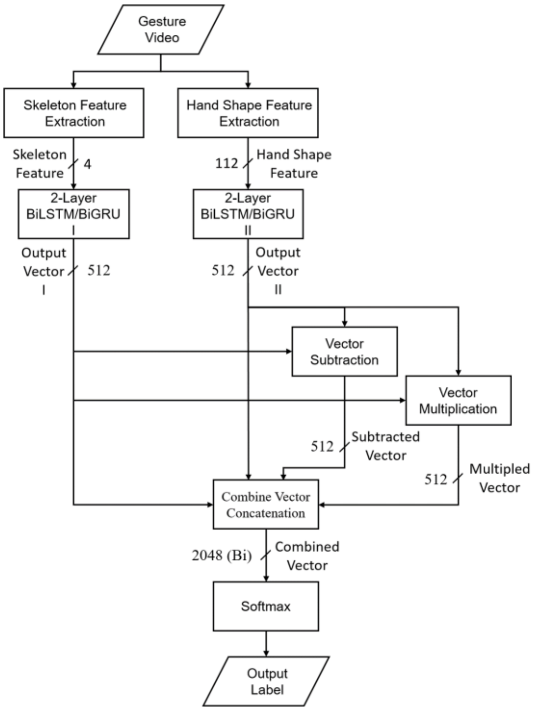


Fig. 9. Conneau Feature Vector Concatenation Model

into three stages: Preparation, Concatenation, and Recognition stage, as seen in Fig 10.

The Preparation stage begins by extracting both feature sets from each video frame. The per-frame features are preprocessed and chained into 19-frame-long-per-word sequences, in line with prior works on this topic [16]–[19]. Nineteen frames' worth of features represents each word in the sentence, and the frames are chosen to be those that contain the most information about that word. The last step is dividing the data into five-folds, whereby four folds are used for training and one fold is used for testing.

The first step in the Concatenation stage is training Model-I with the skeleton feature set and training Model-II with the handshape feature set. In line with [16]–[19], Models-I and II have a 256-node size hidden layer. Since they are bidirectional, and the input is of size 19 * 1, they end up outputting a 512 * 19 feature tensor each. The final step is concatenating the (512 * 19) output from Model-I and the (512 * 19) output from Model-II, which will result in a (1024 * 19) concatenated tensor.

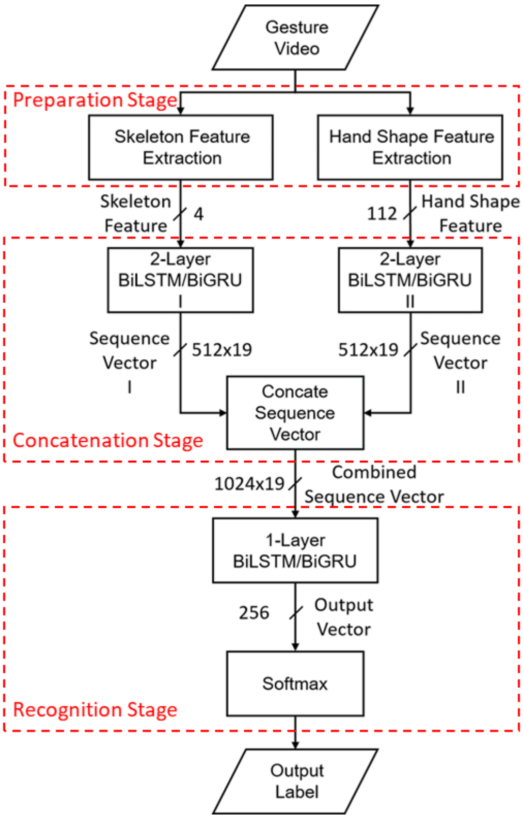


Fig. 10. Sequence Feature Vector Concatenation Model

The Recognition stage uses the concatenated tensor as input to the Model-III. Model-III outputs a 256 * 1 feature tensor. This penultimate feature tensor is then fed into a softmax to calculate the likelihood of each word.

Table 3 illustrates that both the 2-layer bi-directional GRU and LSTM work best with this data, which is why the architecture is chosen for RNN Models I and II. Conversely, Model III is a 1-layer Bi-LSTM and Bi-GRU. The proposed experiment results can be seen in Table 7 below.

Table 7 shows that the best performance was obtained using a 2-layer BiLSTM for Models-I and II and a 1-layer BiLSTM for Model-III. Concatenating these features results in a maximum accuracy of 88.016%, which improves the 82.849% obtained from using the skeleton features only. It is also an improvement from the 55.530%

Table 7. RNN Prediction Accuracies by using Sequence Feature Vector Concatenation Model

Feature	RNN Model		Fold					
	I-II	III	1	2	3	4	5	Average
Sequence Feature Vector Concatenation	**2-Layer BiLSTM**	**1-Layer BiLSTM**	88.062%	88.007%	88.138%	87.504%	**88.370%**	**88.016%**
		1-Layer BiGRU	**88.336%**	86.968%	87.231%	85.561%	86.827%	86.985%
	2-Layer BiGRU	**1-Layer BiLSTM**	**86.998%**	86.298%	86.828%	86.298%	86.003%	86.485%
		1-Layer BiGRU	86.106%	**86.265%**	85.719%	85.863%	85.660%	85.923%

Table 8. Summary of Accuracy versus Model and Concatenation Techniques

Model	Fold	Skeleton	Hand	SFC	FVC	CFVC	SFVC-BiLSTM	SFVC-BiGRU
Bi-LSTM	1	0.8329	0.5983	0.4127	0.8391	0.8590	0.8806	0.8834
	2	0.8372	0.5132	0.4127	0.8338	0.8513	0.8801	0.8697
	3	0.8353	0.4970	0.4983	0.8474	0.8525	0.8814	0.8723
	4	0.8171	0.5749	0.5531	0.8328	0.8623	0.8750	0.8556
	5	0.8199	0.5931	0.6041	0.8329	0.8473	0.8837	0.8683
	Fold	**Skeleton**	**Hand**	**SFC**	**FVC**	**CFVC**	**SFVC-BiLSTM**	**SFVC-BiGRU**
Bi-GRU	1	0.8322	0.4415	0.5396	0.8271	0.8237	0.8700	0.8611
	2	0.8231	0.4975	0.4623	0.8379	0.8204	0.8630	0.8626
	3	0.8380	0.5336	0.1969	0.8380	0.8474	0.8683	0.8572
	4	0.8178	0.3591	0.5226	0.8168	0.8345	0.8630	0.8586
	5	0.8182	0.5029	0.5108	0.8326	0.8398	0.8600	0.8566

obtained independently using the hand shape features. features. These results are the best we have obtained so far.

5. Experimental Result and Analysis

This section discusses all experimental results, both in terms of accuracy and the time required for training and testing.

5.1 Accuracy Comparison

Table 8 summarizes the overall accuracy obtained from the skeleton, hand shape features, and the four techniques combining the two features. Based on Table 8, a two-way ANOVA was conducted to compare the effects of each concatenation technique and the performance of 2-layer bi-directional LSTM and GRU. The concatenation techniques' effects were statistically significant on accuracy. It was explained by the F-value = 122.01 (> $F_{critical}$ 4.28), and P-value = 5.31E-06 (<0.05). The recognition ability of 2-layer bi-directional LSTM and GRU was statistically not significant as the F-value = 4.83 (< $F_{critical}$ 5.99) and the P-value = 0.0702 (>0.05).

5.2. Time Comparison

In addition to accuracy, we also compare the time it takes to reach convergence during training (Table 9) and the average time it takes to recognize each label in the testing data, as shown in Table 10. Table 11 compares testing time versus the accuracy for the best accuracy in each feature extraction technique. The Sequence Feature Vector Concatenation began to converge at epoch 45, the fastest among all the models. The inference time for this model is 0.03 seconds, which is the longest inference time, but the difference is only 0.02 seconds with the fastest inference time of 0.01, which is obtained from the skeleton only feature. The difference in inference time of 0.02 seconds is valuable for better accuracy

6. Conclusion and future research

SIBI gestures can be divided into four categories: gestures that have a unique skeleton

Table 9. Training Time Comparison

Fold	Training Time (in seconds)					
	Skeleton (2-Layer BiGRU)	**Hand Shape (2-Layer BiGRU)**	**Simple Feature Concatenation Model (2-Layer BiLSTM)**	**Feature Vector Concatenation (2-Layer BiLSTM)**	**Conneau Feature Vector Concatenation (2-Layer BiLSTM)**	**Sequence Feature Vector Concatenation (2-Layer BiLSTM + 1-Layer BiLSTM)**
1	10380	12471	12461	25466	25526	36070
2	10347	12376	12400	25340	25538	35745
3	10348	12385	12405	25444	25568	35872
4	10357	12352	12408	25416	25569	35848
5	10545	12471	12704	25605	25737	36146
Average	**10395**	**12411**	**12476**	**25454**	**25588**	**35936**

Table 10. Testing Time Comparison

Fold	Test Time Each Label (in seconds)					
	Skeleton (2-Layer BiGRU)	**Hand Shape (2-Layer BiGRU)**	**Simple Feature Concatenation Model (2-Layer BiLSTM)**	**Feature Vector Concatenation (2-Layer BiLSTM)**	**Conneau Feature Vector Concatenation (2-Layer BiLSTM)**	**Sequence Feature Vector Concatenation (2-Layer BiLSTM + 1-Layer BiLSTM)**
1	0.01084	0.01336	0.01341	0.02535	0.02551	0.03139
2	0.01083	0.01340	0.01342	0.02542	0.02549	0.03122
3	0.01089	0.01339	0.01339	0.02546	0.02558	0.03051
4	0.01087	0.01342	0.01342	0.02543	0.02549	0.03098
5	0.01086	0.01339	0.01345	0.02547	0.02554	0.03013
Average	**0.01086**	**0.01339**	**0.01342**	**0.02542**	**0.02552**	**0.03085**

Table 11. Testing Time versus Accuracy Comparison

Feature	Total Epoch to Convergence	Average Inference Time per Word in Test Set (in seconds)	Best Accuracy
Skeleton (2-Layer BiGRU)	94	**0.0108578**	0.83803763
Hand Shape (2-Layer BiGRU)	703	0.013392034	0.5336
Simple Feature Concatenation Model (2-Layer BiLSTM)	310	0.013416822	0.60411664
Feature Vector Concatenation (2-Layer BiLSTM)	266	0.025424432	0.84744624
Conneau Feature Vector Concatenation (2-Layer BiLSTM)	367	0.025520799	0.86231156
Sequence Feature Vector Concatenation (2-Layer BiLSTM + 1-Layer BiLSTM)	**45**	0.030845543	**0.88370497**

movement, gestures that have uniques hand shapes, gestures that have the same skeleton movement but have different hand shapes, and gestures that have the same hand shape but different skeletal movements.

This study uses two features to capture the characteristics of SIBI gestures. The first feature is the skeleton feature which is data on the angle between the shoulder and the upper arm, and the angle between the upper arm and the forearm. The second feature is the hand shape feature which consists of the area of the hand blob, the centroid of the hand, and the points of the contour of the hand shape as a result of the Elliptical Fourier Descriptor.

To recognize the four groups of SIBI gestures, combining skeleton features and hand shape features is required. The feature combination techniques that have been tried are: Simple Feature concatenation, Feature Vector concatenation, Conneau Feature Vector Concatenation, and Sequence Feature Vector Concatenation.

The average highest SIBI gesture recognition accuracy was achieved by Sequence Feature Vector Concatenation of 88.016%, followed by Conneau Feature Vector Concatenation of 85.448%, Feature Vector concatenation 83.723%, model with Skeleton feature only 82.849%, model with Handshape feature only by 55.530%. and the lowest by Simple Feature Concatenation 49.618%.

The experiment results show that the combined features of the whole gesture sequence can better distinguish one gesture from another in SIBI than the combined features of each gesture frame. The fastest convergence was achieved by Sequence Feature Vector Concatenation at epoch 45. This proves that the combined features of one sequence can quickly distinguish one gesture from another.

The best inference time is achieved by the model with the Skeleton feature only of 0.01086 seconds. We can conclude that the $\theta_1 - \theta_4$ angles present in the skeleton features can quickly distinguish one gesture from another.

The RNN models that give the best performance are 2-layer Bidirectional LSTM and 2-layer Bi-directional GRU. The location of the differentiating gesture components for SIBI's gestures is why the Bidirectional RNN can recognize almost all types of hand movement in SIBI.

The concatenation techniques' effects were statistically significant on accuracy. It was explained by the F-value = 122.01 (> Fcritical 4.28), and P-value = 5.31E-06 (<0.05). The recognition ability of 2-layer bi-directional LSTM and GRU was statistically not significant as the F-value = 4.83 (< Fcritical 5.99) and the P-value = 0.0702 (>0.05).

Sequence Feature Vector Concatenation can improve the ability of the skeleton and hand shape features to recognize SIBI signals.

In the future, we will continue to improve SIBI's gesture recognition ability by trying other hand shape features that have better capabilities than the handshape feature in this study. We also plan to use Convolutional Neural Networks (CNN) like ResNet, MobileNet.

Acknowledgement

This work is supported by the Ministry of Research and Technology Research Grant PTUPT, Number NKB-306/UN2.RST/HKP.05.00/2020. This support is gratefully received and acknowledged. The authors also wish to thank Lim Yohanes Stefanus and M. I. Mas for the final proofreading.

References

[1] *Kamus Sistem Isyarat Bahasa Indonesia*. Jakarta: Direktorat Pendidikan Luar Biasa, 2002.

[2] S. Siswomartono, *Cara Mudah Belajar SIBI (Sistem Isyarat Bahasa Indonesia)*. Jakarta: Federasi Nasional untuk Kesejahteraan Tunarungu Indonesia, 2007.

[3] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "SIGN LANGUAGE RECOGNITION BASED ON HAND AND

BODY SKELETAL DATA," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2018, pp. 1–4.

[4] S. Pramada and A. Vaidya, "Intelligent Sign Language Recognition Using Image Processing," *IOSR J. Eng.*, vol. 3, pp. 45–51, 2013.

[5] A. S. Nikam and A. G. Ambekar, "Sign language recognition using image based hand gesture recognition techniques," in *International Conference on Green Engineering and Technologies*, 2016, pp. 1–5.

[6] E. Rakun, M. Adriani, I. W. Wiprayoga, K. Danniswara, and A. Tjandra, "Combining depth image and skeleton data from Kinect for recognizing words in the sign system for Indonesian language (SIBI [Sistem Isyarat Bahasa Indonesia])," in *IEEE International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2013, pp. 387–392.

[7] E. Rakun, A. M. Arymurthy, L. Y. Stefanus, A. F. Wicaksono, and I. W. W. Wisesa, "Recognition of Sign Language System for Indonesian Language using Long Short-Term Memory Neural Networks," *Adv. Sci. Lett.*, vol. 4, no. 2, pp. 400–407, 2016.

[8] M. A. Bencherif, M. Algabri, M. A. Mekhtiche, M. Faisal, and M. Alsulaiman, "Arabic Sign Language Recognition System Using 2D Hands and Body Skeleton Data," *IEEE Access*, vol. 9, pp. 59612–59627, 2021.

[9] J. Huang, W. Zhou, H. Li, and W. Li, "SIGN LANGUAGE RECOGNITION USING 3D CONVOLUTIONAL NEURAL NETWORKS Jie Huang , Wengang Zhou , Houqiang Li , and Weiping Li University of Science and Technology of China , Hefei , China," *IEEE Int. Conf. Multimed. Expo*, pp. 1–6, 2015.

[10] M. Elpeltagy, M. Abdelwahab, M. E. Hussein, A. Shoukry, A. Shoala, and M. Galal, "Multi-modality based Arabic sign language recognition," *Inst. Eng. Technol. Comput. Vis.*, vol. 12, 2018.

[11] A. Conneau and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," in *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[12] D. A. Kumar, S. Member, A. S. C. S. Sastry, P. V. V Kishore, S. Member, E. K. Kumar, S. Member, M. T. K. Kumar, and S. Member, "S3DRGF : Spatial 3-D Relational Geometric Features for 3-D Sign Language Representation and Recognition," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 169–173, 2019.

[13] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel Transformers for Multi-articulatory Sign Language Translation," in *16th European Conference on Computer Vision*, 2020.

[14] 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.

[15] E. Rakun, "Pengenalan Komponen Imbuhan dan Kata Dasar pada Isyarat Kata Berimbuhan dalam SIBI(Sistem Isyarat Bahasa Indonesia) dengan Menggunakan Probabilistic Graphical Models," University of Indonesia, 2017.

[16] K. Halim and E. Rakun, "Sign Language System for Bahasa Indonesia (Known as SIBI) Recognizer using TensorFlow and Long Short-Term Memory," in *International Conference on Advanced Computer Science and Information Systems (ICACSIS*, 2018, pp. 403–407.

[17] M. Harits, E. Rakun, and D. Hardianto, "Feature Extraction from Smartphone Images by Using Elliptical Fourier Descriptor, Centroid and Area for Recognizing Indonesian Sign Language SIBI (Sistem Isyarat Bahasa Indonesia)," in *2nd International Conference on Intelligent Autonomous Systems*, 2019.

[18] A. Aulia, E. Rakun, and D. Hardianto, "Human Skeleton Feature Extraction from 2-Dimensional Video of Indonesian Language Sign System (SIBI [Sistem Isyarat Bahasa Indonesia]) Gestureso Title," in *ACM Conference Proceedings*, 2019.

[19] K. Anggraini, E. Rakun, and L. Y. Stefanus, "Recognizing the Components of Inflectional Word Gestures in Indonesian Sign System known as SIBI (Sistem Isyarat Bahasa Indonesia) by using Lip Motion," in *International Conference on Electrical Engineering and Informatics (ICEEI)*, 2019.