American Sign Language Words Recognition using Spatio-Temporal Prosodic and Angle Features: A sequential learning approach

SUNUSI BALA ABDULLAHI1,2, (Member, IEEE), KOSIN CHAMNONTAI3, (SENIOR MEMBER, IEEE)

1Department of Computer Engineering, Faculty of Engineering, King Mongkut’s University of Technology Thonburi, 126 Pracha-Uthit Road, Bang Mod, Thung Khru, Bangkok 10140, Thailand. (e-mail: sbabdullahi@ieee.org)
2Force Criminal Investigation and Intelligence Department, Nigeria Police, 900211, Louis Edet House Force Headquartes Garki.Shehu Shagari Way, Abuja
3Department of Electronics and Telecommunication Engineering, Faculty of Engineering, King Mongkut’s University of Technology Thonburi, 126 Pracha-Uthit Road, Bang Mod, Thung Khru, Bangkok 10140, Thailand. (e-mail: kosin.cha@kmutt.ac.th)

Corresponding author:(e-mail: kosin.cha@kmutt.ac.th).

This work is financially supported by Petchra Pra Jom Klao Ph.D. Research Scholarship (Grant No. 45/2562) from King Mongkut’s University of Technology Thonburi

ABSTRACT Most of the available American Sign Language (ASL) words share similar characteristics. These characteristics are usually during sign trajectory which yields similarity issues and hinders ubiquitous application. However, recognition of similar ASL words confused translation algorithms, which lead to misclassification. In this paper, based on fast fisher vector (FFV) and bi-directional Long-Short Term memory (Bi-LSTM) method, a large database of dynamic sign words recognition algorithm called bidirectional long-short term memory-fast fisher vector (FFV-Bi-LSTM) is designed. This algorithm is designed to train 3D hand skeletal information of motion and orientation angle features learned from the leap motion controller (LMC). Each bulk features in the 3D video frame is concatenated together and represented as a high-dimensional vector using FFV encoding. Evaluation results demonstrate that the FFV-Bi-LSTM algorithm is suitable for accurately recognizing dynamic ASL words on basis of prosodic and angle cues. Furthermore, comparison results demonstrate that FFV-Bi-LSTM can provide better recognition accuracy of 98% and 91.002% for randomly selected ASL dictionary and 10 pairs of similar ASL words, in leave-one-subject-out cross-validation on the constructed dataset. The performance of our FFV-Bi-LSTM is further evaluated on ASL data set, leap motion dynamic hand gestures data set (LMDHG), and Semaphoric hand gestures contained in the Shape Retrieval Contest (SHREC) dataset. We improve the accuracy of the ASL data set, LMDHG, and SHREC data sets by 2%, 2%, and 3.19% respectively.

INDEX TERMS American Sign Language, Deep learning, Fast fisher vector, Hand gesture recognition, Leap motion controller, Orientation angles, Spatio-temporal Sequence, Ubiquitous computing.

I. INTRODUCTION

THE incredible attention in human-computer interaction (HCI) makes human hands the most natural and efficient medium to express intentions for daily interaction activities [1]. It leads to the development of numerous HCI systems such as sign language recognition, robotics, medical diagnostics, among others. Deaf are generally dependent on sign language to participate in the real world. World Federation of the Deaf put figures around three hundred active natural sign languages across the globe [2]. American Sign Language (ASL) is one of the famous sign languages with unwritten grammar characterized by hand motions, and sometimes facial/body signs [3]. This language involves constructing very complex grammatical structures, using dynamic word gestures. The dynamic word gestures are most crucial constructing blocks during ASL sentence development and facilitating expressive communication. ASL comprises over ten thousand dynamic word gestures with approximately 65% and 35% represented
by sign words and finger-spelled words respectively [4]. Sign words remain the common means for the deaf to express themselves. Therefore, these words are indispensable for daily deaf communication. It is imperative to mention that majority of the available ASL words comprised of similar gestures. Thus, the similarity usually confuses sensing devices and hinders the application of most sensors leading to misclassification. To solve this, Fang et al. [5], proposed DeepASL using leap motion controller (LMC) sensor from backhand view with bi-directional long short term memory (Bi-LSTM). Therefore, Deep Bi-LSTM architectures should have more potential for the dynamic sign language recognition (SLR) [6], [7].

In Avola et al. [6], a similar recent approach where LMC with stack Deep Bi-LSTM network is used as a prediction model on temporal feature descriptors, which represent coordinates of internal hand joints angles and the palm displacement. However, stacking large number of Deep Bi-LSTM units resulted to unsatisfactory recognition accuracy. Motivated by [5], [6], we present 3D Spatio-temporal skeletal hand joint features according to the prosodic model and orientation angle to address misclassification of highly correlated ASL words. These words are difficult to be recognized by learning internal hand joint angles and the palm displacement only, thus, the similar ASL words can be treated as composed by many small orientation variations and prosodic cues. The major difference between the Deep Bi-LSTM in [6] and ours, is that, we trained the Deep Bi-LSTM from encoded fast fisher vector (FFV) information to improve the Deep Bi-LSTM learning and reduce large abstraction. Our contributions are supported by several sign language models [8]–[11]. We make the following contributions:

(i) We introduced orientation angle \( Q_n \) and prosodic \( \mu \) features to discriminate similarity between ASL words from 3D skeletal hand characteristics.

(ii) Developed robust fast fisher vector (FFV) for feature selection and encoding in Deep Bi-LSTM, which requires no large abstraction.

(iii) Hyper-parameters tuning of FFV-Bi-LSTM sequential learning algorithm is conducted using a validation data-driven approach.

(iv) We classified complex gestures using FFV-Bi-LSTM that are critical to recognize by conventional Deep Bi-LSTM algorithms.

(v) Our method conforms with the existing results in numerous examples, even with a limited number of data set, static and dynamic hand gestures.

The remainder of this article is as follows: Section II introduces related works. Section III provides problem analysis, mathematical hand gesture models, spatio-temporal feature extraction, data correction and normalization, FFV encoding, and FFV-Bi-LSTM). The recognition phase is proposed in Section IV-A2. Section IV provide details of experimental analysis and evaluation. Discussion is proposed in Section V. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

From the existing works, we can further subgroup available SLR systems into four groups as shown in Table 1. The first group addressed SLR sensing using a contact-based system, which is further sub-divided into two classes namely: wearable systems [12]–[16], which are very unnatural and prone to misclassification and radio frequency system (RF) [17]–[19] more natural and address intrusion, however, these systems are restricted to high internet access and interference. The emergence of digital cameras and camera stereo gave birth to the vision-based SLR, forming the second group [20], [21], [21]–[30] are natural, however, the camera systems suffer complex segmentation. Sensors such as optical sensors, flex sensors, accelerometers, etc. [16], [31]–[34] require no segmentation and good accuracy. However, they are very expensive, invasive, unnatural, and needs calibration set, as shown in Table 1. Therefore, recent papers track dynamic sign words using active imaging devices such as LMC [1], [5], [35], MS Kinect [36] and Orbbec Astra which are portable, requires no complex segmentation, no calibration, inexpensive, mobile, and provides 3D information. This formed an active image sensor-based group four. The summary of some of the available recognition methods are illustrated in Table 2.

III. MATERIALS AND METHODS

In this section, our approach for addressing the misclassification problem consists of the following process: Problem analysis, mathematical hand gesture models, spatial and temporal feature extraction, data correction and normalization, FFV encoding, and lastly FF-Bi-LSTM algorithm. This procedure is illustrated in Fig. 1.

A. PROBLEM ANALYSIS

To solve misclassification, authors in [6] utilizes skeletal joint sequence of hand displacements and internal angles as their feature vector. However, these features are insufficient to recognize most ASL words, especially similar ASL words in Figs. (2)-(3). It is found that the differences among these ASL words happen more at hand orientation as shown in Figs. 2(a), (e), (f) and 3(a) and (d). However, small motion at wrist generate large variation angles \( (\Delta \gamma) \). To analyze hand orientation, there is need to investigate prosodic model as described in [10]. The Prosodic model is built from Inherent and prosodic cues to form a lexeme at the root node. Inherent cues comprised of handshape, location and orientation. Prosodic cues are motion (movement cues) features. This is the reason why motion features are known as prosodic features, as shown in Fig. 5. Thus, prosodic cues are mathematically represented to mimic hand joint motion.

B. MATHEMATICAL HAND GESTURE MODELS

Hand joints are represented in Fig. 4 according to 3D coordinates \( X, Y, \) and \( Z \) axes, which set origin at wrist position. The distance \( X_{j,k} \) between positions \( j \) and \( k \) gives the relationship
Table 1: SLR according to capturing modalities

| Algorithm name | Brief methodology | Highlights | Limitations |
|----------------|-------------------|------------|------------|
| Talking Hands [12] | Distance function + glove + smartphone speech synthesizer | Real-time via scenario translation | Intrusive, bulk and unnatural |
| MyoSign [13] | Myo arm band signal Multi-CNN + BiLSTM + CTC | No temporal segmentation Hand shape + trajectory | Real-time + 3D printed humanoid region growing technique occlusion |
| Seyedarahi et. Al. [15] | White glove signal HMM + Gaussian | Multi-task learning Hand kinematics + SL analysis | 3D printed glove + humanoid cumbersome |
| Data glove [16] | Data glove + IMU + TOF + FSR | Real-time + 3D printed humanoid Hand kinematics + SL analysis | Gesture recognition consuming + invasive |
| WiSign [19] | WiSign’s signal + CSI + PSD DB of 35000 manual units | High internet access Multimodal gestures | Non-ubiquitous Internet access + interference |
| SignFI [17] | WiSign’s signal + CSI | CSI measurements Multimodal gestures | Non-ubiquitous Internet access + interference |
| WiGest [18] | WiSign’s signal + CSI | No gesture learning Ubiquitous system | High internet access + interference |

Digital camera and camera stereo SLR methods

| Algorithm name | Brief methodology | Highlights | Limitations |
|----------------|-------------------|------------|------------|
| ArSLRS [23] | RGB videos + YOLOv3 color space | YOLOv3 segmentation multimodal fusion | Segmentation complexity Complex environment + skin effect |
| Xue et al. [24] | Voting strategy + deep forest | Semantic consideration skeleton projection | Complex learning |
| JDTD and JATD [20] | Multimodal RGB + OpenPose | two stream of CNN | | |
| DNN [21] | 3D motion camera information | RGB End-to-end learning RGB | Skin effect |
| Ranstoo et al. [22] | Multimodal RGB + Depth | 3D multimodal fusion | Segmentation complexity 3D multimodal fusion |
| ASLNN [25] | LSTM + CNN | Optical + scene flow LSL + CHA + ANN | 2D projection looks alike |
| Tran et al. [27, 28] | Smartphone-based capturing | Human Signal intelligibility model | | |
| Au-Swipe Gesture [29] | Smartphone-based capturing | Ubiquitous SLR + OpenCV | | |
| Settle SLRs [30] | Smartphone-based + DCT + PCA | Self-based capturing Not applicable while walking | | |
| Lim et al. [26] | MDC + Euclidean distance | Gaussian pre-filtering SLR from iconic structure | | |
| Dicta-Sign-LSF-v2 [37] | Hardware filter + HEI + GEI + CNN | DB of 35000 manual units | Multimodality |

Sensor-based SLR methods

| Algorithm name | Brief methodology | Highlights | Limitations |
|----------------|-------------------|------------|------------|
| Jinhareenport [32] | ILT + LDA + k-NN | sensor-based capturing | | |
| Chu et al. [33] | Residual PairNets + MAP | Accelerometers + gyroscopes | Calibration + cumbersome |
| Stretchable e-skin [34] | Backhand-view based capturing | multiple sensors | Low accuracy |
| | | | cumbersome |
| | | | Pervasive + trial and error |
| | | | Unnatural |

Active imaging device SLR methods

| Algorithm name | Brief methodology | Highlights | Limitations |
|----------------|-------------------|------------|------------|
| Kumar et al. [36] | Kinect skeleton coordinate + HMM | Real-time position invariant system | hard learning + Limited FoV |
| Auerljus et al. [35] | LMC + HMM | Dynamic hand gestures | Limited representation ability |
| DeepASI [35] | LMC + HMM | Dynamic hand gestures | Limited representation ability |
| TheRusLan [38] | Motion capture based LMC + HBRNN | Ubiquitous + Real-time SLR | SL database design |

Table 2: Sign language recognition methods

| Algorithm name | Brief methodology | Highlights | Limitations |
|----------------|-------------------|------------|------------|
| Kasuluk and Napela [39] | SVM + evolutionary strategy | vector differences between shapes | separate features learning low representation ability |
| Alimuzare and Al-Nuaimi [40] | SVM with SVM + RF + KNN | low representation ability | separate features extraction learning small variation fails |
| GMM-HMM [41] | WLR + GMM-HMM | key frames denotes hidden states | separate features extraction learning small variation fails |
| Tornay et al. [42] | continuous + HMM | joint region recognition of AU | limited learning ability |
| da Silva et al. [43] | CNN + CNN-LSTM | Video analysis of AU’s with FACS | Segmentation complexity |
| Polat and Sarafaral [44] | UTD + KNN | RGB videos from OP and HD | Information looks alike |
| Pareti et al. [45] | Attention-based CNN | Pose and shape KWS of OP | decrease in accuracy due to ED |
| KWS [46] | KWS + end-to-end CNN | OpenPose from RGB videos | Segmentation |
| De Coster et al. [47] | OpenPose + MTNs | Relationship between signs | constrained condition of fusion model |
| Zhang et al. [48] | OpenPose + MTNs | OpenPose from RGB videos | Segmentation complexity |
| Yuan et al. [49] | YOLOv3-STN + Pica-Bayes | OpenPose from RGB videos | | |
| Mujiash et al. [50] | YOLOv3 + DarkNet-53 | OpenPose from RGB videos | Segmentation complexity |
| LSTMM+CHMM [51] | LSTMM + CHMM + CNN | On-the-fly + segmentation of static gestures | separate features learning low representation ability |
| Averla et al. [52] | Multi-stuck LSTM | Detection of skeletal joints | separate features learning low representation ability |
| Bull et al. [52] | Multi-stuck LSTM | Detection of skeletal joints | separate features learning low representation ability |
| Borg et al. [53] | OpenPose + factorization | substrates model-based SLR | segmentation issues |
| NRSIM + Bi-RNN-CAC | | | | |
| MEDIAP-SKEL 2D [54] | OpenPose + concodancer + CNN | preserve phonology meaning | Bi-RNN sometimes may lead to memory explosion especially when the scene changes |
| Kaczmarek and Filhol [55] | Bra + Elan + CAT | List of categorical objects in SL | | |
| Mukesh et al. [56] | OpenPose + Logistic regression | non-deep learning method | | |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Figure 1: Flow chart of the proposed method

Figure 2: Highly correlated double hand ASL words **(Good)** and **(Bad)**: In Figs. (a)-(f) shows corresponding 3D feature representations of prosodic model. Their corresponding angle domain waveform is shown in (a.) and (d.). Corresponding 3D hand joints motion waveform is represented in (b.) and (e.). Pictures (c.) and (f.) shows corresponding hand shape waveform.
Figure 3: Highly correlated single hand ASL words (Hey) and (Child): In pictures (a)-(f.) shows corresponding 3D feature representations of prosodic model. Their distinct corresponding 3D angle waveform is shown in (a.) and (d.). The Corresponding 3D hand joints motion waveform is represented in (b.) and (e.). In pictures (c.) and (f.) shows corresponding 3D hand shape waveform.

Figure 4: Skeleton hand joints definitions

between finger joints and fingertips \( (Z_{j,k,l}) \), equivalently written as

\[
Z_{j,k,l} = [t's/po(j3), tj/tk(j3, tj), t's/j's(t, j)].
\]

(1)

Where \( t's/po, tj/tk, \) and \( t's/j's \) denotes all fingertips to palm, fingertip to fingertip, and fingertip to fingertip to joint ratios, respectively. Then, the prosodic features \( \mu \) of finger joints motion \( M^f(n) \) per each frame \( f \) can be coined as \( \nu^f(n) \), where \( n \) denotes number of sequence per each frame. Thus,

\[
\mu = \{M^f + \nu^f + Z^f\}.
\]

(2)

Figure 5: Similar ASL words using single hand according to prosodic model

Similarly, the chosen mathematical representation for hand orientation angle about motion axis \( Y_R \) was a Right-hand rule, which can be obtained using cross-product as follows

\[
Y_R = \frac{Z_R \times X_L}{|Z_R \times X_L|}.
\]

(3)

Thus, angle between \( Z_R \) and \( X_L \) is denoted as \( \varphi \). Similarly, hand internal angles \( b \) can be obtained according to finger joint angles as shown in Fig. 4. Finally, hand orientation angles can be put together as angular feature vector \( Q \), defined as

\[
Q = \{\varphi + a + b\}.
\]

(4)

Therefore, extracted features according to formulations in Eqs. (1)-(4) are fused through simple vector concatenation equivalently written as:
where \( v \in \{\text{thumb}, \text{index}, \text{middle}, \text{ring}, \text{little}\} \), and \( \rho \) contains inherent features. Solving this model, a state-of-the-art Deep FFV-Bi-LSTM algorithm is adopted.

### C. SPATIO-TEMPORAL FEATURE EXTRACTION

Spatio-temporal features are basically defined by given frame length \( F \) of sequence matrix

\[
L = [M_1, M_2, \ldots, M_F]
\]

(6)

Each matrix \( M_t \in L \) consists of skeletal measurements at time-step \( t \).

Thus, spatial information is obtained by setting a threshold value among successive video frames, as given in Eq. (9).

\[
\text{maximum velocity (peak velocity), as illustrated in frames } \frac{\lambda}{\text{time-step}} \text{ of } M_l \geq 45\%.
\]

Moreover, temporal features are the hand coordinates of all finger joints, tips of hand, palm center, and wrist center, which generates approximated 3D coordinates of 22 poses. The pose is distinguished by velocity, that is \( \frac{\lambda}{\text{time-step}} \geq 45\% \). Therefore, spatial information is obtained by setting a threshold \( \lambda \) as shown in Eq. (8), which is handled by Savitzky-Golay smoothing filter. Thus, generative model parameters \( \theta \) and covariance matrix \( \sigma \) of the Gaussian respectively; \( k \) denotes the number of Gaussian distributions in the mixture model, which is learned together with the features vector as follows:

\[
\theta = \{w_k, \mu_k, \sigma_k, \mu_k, \sigma_k, \ldots, cov_k : k = 1 \cdots K\}.
\]

To apply FFV features, let \( \lambda = \{\lambda_t : t = 1 \cdots T\} \) be the set of \( T \) local information in Eq. (8), thus, generative procedures \( \lambda \) of the whole feature vectors are formulated as follows

\[
H_\theta(\lambda) = \frac{1}{t} \sum_{k=1}^{K} (\lambda_t ; \mu_k, \sigma_k)w_k,
\]

(10)

Also, FFV matrix can be obtained as follows:

\[
\mathbf{x}_\lambda = [\nabla_{\theta} \log \mu_\theta(T) \nabla_{\theta} \log \mu_\theta(T)]^T.
\]

(11)

Similarly, \( \mathbf{x} \) is finally obtained from fused partial derivatives through GMM parameters

\[
\mathbf{x}^t = [G_{\mu,1}, \mu^t_{s,1}, \ldots G_{\mu,k}, G_{s,k}^t],
\]

(12)

Where \( H_\theta, 1/v, \nabla_{\theta} \log(\cdot) \) denote generative model parameters, normalized values, and log-likelihood gradient. The \( \theta \) are discover from training features via expectation maximization (EM) strategy. Gradients are computed according to mean vector \( \mu_f \) and standard deviation \( (s_k) \) of the \( f \)th Gaussian in Eq. (12).
Figure 6: Data correction: A. shows original average skeletal hand video frames, and B. represents smoothed and corrected frames information using weighted linear regression.

Figure 7: 3D keypoints generation with Fast Fisher vector transformation.

F. FAST-FISHER-BI-LSTM (FFV-BI-LSTM)

A Combination of FVs and deep neural networks was already considered [59]. But FFV (GMM with diagonal covariances) has not been considered in Deep Bi-LSTM for SLR [4], [5], [51], [60]–[62]. Features encoded by FFV are concatenated numerically using three-stacked Bi-LSTM layers as shown in Fig. 8. Basically, each Bi-LSTM layer evaluate FFV encoding, dimension reduction, spatial stacking, and $L_2$ normalization throughout Gaussians and $\lambda$ as follows:

$$
O_{f,n} = \sigma [V_{h,o}^f \overrightarrow{h}_o f, Q_{f,n} + V_{h,o}^f \overrightarrow{h}_f f, Q_{f,n} + V_{h,o}^f \overrightarrow{h}_f f, \mu_{f,n} + + V_{h,o}^f \overrightarrow{h}_f f, \rho_{f,n} + V_{h,o}^f \overrightarrow{h}_f f, \rho_{f,n} + + V_{h,o}^f \overrightarrow{h}_f f, L_{f,n} + + V_{h,o}^f \overrightarrow{h}_f f, L_{f,n} + d_o]$$

(13)

where $\sigma$, $V_{h,o}$, $h_f$, $n$, $Q$, $\mu$, $\rho$, and $L$ denotes logistic sigmoid function, weight matrices, hand index, angle, motion, shape, and spatial features, and $d_o$ denotes bias. Where $\overrightarrow{h}_o$ and $\overrightarrow{h}_f$ denotes forward hidden and cell state vectors. $\overrightarrow{h}_o$ and $\overrightarrow{h}_f$ denotes previous hidden and cell state vectors.

IV. EXPERIMENTAL ANALYSIS AND EVALUATION

A. EXPERIMENT

We evaluates the FFV-Bi-LSTM recognition algorithm using spatial-temporal prosodic and angle features in three cases.

The first, second and third case adopt skeletal video sequence recognition from our proposed dataset, ASL dataset in [6], and public data sets [6], [63], [64] with FFV-Bi-LSTM. The proposed set up is illustrated in Fig. 10, where a Leap motion controller (LMC) is employed at the signer’s chest to capture 3D skeletal hand joints information from backhand view. This is to enable the natural mobility of the signer. The testing environment is provided in Fig. 12 and the set up values is given in Table 3.

Table 3: Simulation environment

| Systems                  | Requirements                      |
|--------------------------|-----------------------------------|
| Personal Computer        | Dell G3 15 Gaming                |
|                          | CPU: Intel Core i7-9th Gen       |
|                          | Memory Size: 8GB DDR4            |
|                          | Hard Disk Drive: 500 GB          |
| Leap Motion controller   | Frame rate: 120 fps              |
|                          | Weight: 32g                      |
|                          | Infrared camera: 2 x 640 x 240   |
|                          | Range: 80 cm                     |
|                          | FOV: 150 x 120 degrees           |
| Video                    | 30 fps                           |
| Signers                  | 10 persons                       |
| Settings                 | frequency: 10 times per word     |

1) Data sets

In our new datasets, we employed and trained 10 voluntarily hearing ability people to perform 57 randomly selected
ASL words of both single and double hand information. All signers perform the sign while walking and standing. Each signer performs all 57 ASL words, ten (10) times. We have collected 10 pairs of similar ASL words out of 57 ASL words in the dictionary. The selected words belong to frequently used daily first 100 ASL words. Some examples of our datasets are given in Fig. 5. The dataset is partitioned into training and testing using different types of signers (signer-independence). The selected features have undergone various tests to ensure effectiveness. We further evaluate our method on Semaphoric hand gestures contained in the Shape Retrieval Contest (SHREC) [64], ASL Data set [6], and Leap motion dynamic hand gestures (LMDHG) [63] Dataset, respectively.

2) Recognition Phase

Our algorithm calls a function InitialTransformWeights name-value pair. Sparse filtering algorithm is implemented in MATLAB using "sparsefilt" function from yael package. The algorithm handles sparse filtering objective function minimum [65]–[67]. We selected average number of GMM components and few number of iteration for effective video features encoding as provided in Algorithm 1. FFV encoding

Figure 8: Architecture of Bi-directional LSTM

Figure 9: Block diagram of needed hardware components

Figure 10: Photo of experimental system
generates synthetic local information of a particular frame, which do not handle possible time correlation between two different encoded frames of the sequences. To fully exploit this information, three Bi-LSTM units are chosen, each unit accommodate seven layers connected with dropout layers of 20% (0.2) deactivation and validated with careful selection of parameters of Table 4. The total output of this layer is added up and normalized by the softmax layer as shown in Fig. 8. The output $O_{ff}$ from Eq. (15) is considered as probability for a given number of ASL word $L$. For a given $O^L_{ff}$ which have $Lth$ sequence from class $E_{L}$, then the predicted ASL word $G$ is obtained from normalized $O_{ff}$ at softmax. ASL word classification is achieved by computing high probability score $p$ from Eq. (14). The final layer is obtained from the following formulations:

$$O = \sum_{f=0}^{F-1} O^f_{ff}$$  \hspace{1cm} (14)

$$O^L = p(E_{[G]} = \frac{e^{O^L}}{\sum_{L=1}^{L-1} e^{O_i}}, L = 1, \cdots , L$$  \hspace{1cm} (15)

We summarize the steps of sequential gesture recognition in details in the following Algorithm 2.

### Algorithm 2 Sequential Feature learning

1. start
2. set $L$ in Eq. (6) \{Video input sequence\}
3. set $V_h$ \{Sequence weight\}
4. set $S$ \{Sequence length\}
5. set $n$ \{Hand index\}
6. set $f$ \{Sequence length\}
7. for each $n \in [0, f - 1]$ do
8. repeat
9. if $n < s - 1$ then
10. Feed $M_n$ and $V_h$ to Bi-LSTM
11. else if
12. $n \leftarrow S - 1$ then
13. Get $M_n$ from Eq. (6) \{Features for Bi-LSTM\}
14. else if
15. stop
16. end if
17. end for
18. compute parameters and recognition metrics
19. until Eq. (14) converge
20. return Eq. (15)
21. end

V. DISCUSSION

Deep Bi-LSTM with 3 units has hard learning because of high abstraction, which lead to low accuracy. However, Deep FFV-Bi-LSTM has flexible computing which lead to an increase of 5% accuracy. Thus, Deep FFV-Bi-LSTM outperforms the conventional Deep Bi-LSTM in [6]. The superior model is number three with four feature vectors, which is chosen for further analysis. Performance evaluation of model 3 using Deep Bi-LSTM and FFV-Bi-LSTM is demonstrated on Tables 11-12. It is proven that each word takes an amount of 2 seconds to be trained. However, the generalization of model takes approximately 1 second to test each word per sequence. Therefore, the standard deviation of 7.091 is achieved from the mean. This means that each score deviates from the mean by 0.0738 points on average. The accuracy of the algorithm and proposed data set is further evaluated using leave-one-subject-out cross-validation. Per-class accuracy is obtained to be 91.002%, with less than 9.0% error which demonstrates that our algorithm has a high probability to recognize ASL words of similar characteristics, as detailed in Table 10. Table 9 depict the recognition performance of leave-one-subject-out cross-validation of the 57 randomly selected ASL words. Therefore, the chosen mathematical model has proven to be a good choice for our idea. It is also shown that the adopted algorithm has a relatively bad generalization to recognize positive results of “Happy”, “Cheap”, and “Jump”. Research findings show that these similar ASL words have similar spatial information and minimum orientation angle variations. One of the major
Table 5: Results comparison on ASL skeletal data set in [6]

| Approach         | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|------------------|--------------|---------------|------------|--------------|
| Avola et al. [6] | 96.4102      | 96.6434       | 96.4102    | 96.3717      |
| Ours FFV-Bi-LSTM | 98.331       | 98.991        | 98.331     | 98.576       |

Table 6: Results comparison on SHREC dataset

| Approach                  | 14 Hand Gestures | 28 Hand Gestures |
|---------------------------|------------------|------------------|
| De Smedt et al. [68]      | 88.62            | 81.9             |
| SHREC’17 Track [64]       | 82.9             | 71.9             |
| Ohn-Bar and Trivedi [69]  | 83.85            | 76.53            |
| HON4D [70]                | 78.53            | 74.03            |
| Devanne et al. [71]       | 79.61            | 62               |
| Avola et al. [6]          | 97.62            | 91.43            |
| STA-Res-TCN [72]          | 94.4             | 90.7             |
| Liu et al. [73]           | 94.88            | 92.26            |
| Ours                      | 97.99            | 92.99            |

Table 7: Results comparison on LMDHG dataset

| Approach                  | F1-Score (%) |
|---------------------------|--------------|
| Boulahia et al. [63]      | 84.78        |
| Lupinetti et al. [74]     | 92.11        |
| Hisham and Hamouda [75]   | 91.2         |
| Ours                      | 93.08        |

VI. CONCLUSION

In this work, we adopted an approach to recognize highly correlated American sign language words. We optimize the accuracy of recorded 3D video skeletal hand joints information, using a WLR algorithm and filter. The final information is encoded using FFV for fine-grained recognition which depends on a few discriminative features. The Features are found potential and interesting for Deep Bi-LSTM recognition. The second contribution in this article includes the design of a new large 3D dynamic hand skeletal ASL data set. We also systematically compare the radius of convergence of limitations of adopting FFV is trial and error strategy while choosing stable GMM components. All procedures for computing GMM are iterative, therefore emphasis must be put in place on a suitable iteration number for the GMM matrix because of its local convergence.
Table 8: Results comparison with hand shape and motion features

| Data set            | Approach | Accuracy (%) | Number of words | Misclassification (%) |
|---------------------|----------|--------------|-----------------|-----------------------|
| Random              | Ours     | 98.6         | 57              | 2                     |
| Highly correlated   |          | 91.002       | 10 pairs        | 9                     |
| DeepASL [5]         | DeepASL  | 94.5         | 56              | 5.5                   |

Figure 13: Confusion Matrix of the entire dataset

Table 9: Scores per recognized correlated ASL words

| S/no. | Class         | Accuracy (%) | Error (%) |
|-------|---------------|--------------|-----------|
| 1     | Child         | 90           | 10        |
| 2     | Eight         | 100          | 0         |
| 3     | Enthused      | 100          | 0         |
| 4     | Excuse        | 100          | 0         |
| 5     | Expensive     | 90           | 10        |
| 6     | Fork          | 90           | 10        |
| 7     | Happy         | 80           | 20        |
| 8     | Hey           | 90           | 10        |
| 9     | Jump          | 80           | 20        |
| 10    | Like          | 90           | 10        |
| 11    | Bad           | 100          | 0         |
| 12    | Angry         | 90           | 10        |
| 13    | Cheep         | 80           | 20        |
| 14    | Money         | 90           | 10        |
| 15    | Hot           | 90           | 10        |
| 16    | Good          | 90           | 10        |
| 17    | Again         | 90           | 10        |
| 18    | Short in height | 90   | 10        |
| 19    | Dance         | 90           | 10        |
| 20    | Read          | 100          | 0         |
| Total |              | 91.002       | 8.998     |

VII. ACKNOWLEDGEMENTS

This work is financially supported by Petchra Pra Jom Klao Ph.D. Research Scholarship (Grant No. 45/2562) from King Mongkut’s University of Technology Thonburi, Bangkok, Thailand. We are also grateful to the anonymous IEEE Access reviewers for their potential reviews and insightful comments.

References

[1] L. Liu and Y. Huai, “Dynamic hand gesture recognition using lmc for flower and plant interaction,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 33, no. 01, p. 1950003, 2019.
[2] W. F. of the Deaf, “With sign language, everyone is included,” 2018.
[3] M. J. Cheok, Z. Omar, and M. H. Jaward, “A review of hand gesture and sign language recognition techniques,” International Journal of Machine Learning and Cybernetics, vol. 10, no. 1, pp. 131–153, 2019.
[4] P. Chophuk and K. Channongthai, “Backhand-view-based continuous-signed-letter recognition using a rewound video sequence and the previous signed-letter information,” IEEE Access, vol. 9, pp. 40187–40197, 2021.
[5] B. Fang, J. Co, and M. Zhang, “Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation,” in Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems, pp. 1–13, 2017.
[6] D. Avola, M. Bernardi, L. Cinque, G. L. Foresti, and C. Massaroni, “Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphore hand gestures,” IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 234–245, 2019.
[7] S. Ameur, A. B. Khalifa, and M. S. Bouhlel, “A novel hybrid bidirectional
Table 10: Computational cost of proposed method

| Data set                | Training time (sec) | Extraction time (sec) | Number of words |
|-------------------------|---------------------|-----------------------|-----------------|
| Random ASL data set     | 115                 | 59                    | 57              |
| Similar ASL data set    | 21                  | 0                     | 10 pairs        |

Table 11: Different features combination for various Deep FFV-Bi-LSTM model comparison

| Epoch | minibatch size | Model combination                  | Iteration | Processing time (Train) | Processing time (Test) | Accuracy (%) | Learning rate |
|-------|----------------|------------------------------------|-----------|-------------------------|------------------------|--------------|---------------|
| 350   | 27             | Shape + Motion                     | 3500      | 571                     | 85                     | 76           | 1.00E-19      |
| 350   | 27             | Shape + Motion + location          | 3500      | 573                     | 199                    | 81.38        | 1.00E-17      |
| 350   | 27             | Shape + Motion + location + angular features | 10500  | 5232                    | 360                    | 88.086       | 1.00E-17      |

Table 12: Different features combination for various Deep FFV-Bi-LSTM model comparison

| Epoch | minibatch size | Model combination                  | Iteration | Processing time (Train) | Processing time (Test) | Accuracy (%) | Learning rate |
|-------|----------------|------------------------------------|-----------|-------------------------|------------------------|--------------|---------------|
| 350   | 27             | Shape + Motion                     | 2500      | 151                     | 47                     | 76           | 1.00E-25      |
| 350   | 27             | Shape + Motion + location          | 2500      | 211                     | 105                    | 83.98        | 1.00E-19      |
| 350   | 27             | Shape + Motion + location + angular features | 3000  | 295                     | 138                    | 91.002       | 1.00E-20      |

unidirectional lstm network for dynamic hand gesture recognition with leap motion,” Entertainment Computing, vol. 35, p. 1003737, 2020.

[8] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech.-Theory Exp., vol. 2008, p. P10008, 2008.

[9] R. Battison, “Lexical borrowing in american sign language,” 1978.

[10] D. Brentari, Sign language phonology, Cambridge University Press, 2019.

[11] R. A. Tennant, M. Gluszak, and M. G. Brown, The American sign language handshape dictionary. Gallaudet University Press, 1998.

[12] T. Persuoli, D. Corona, M. L. Cordaini, and A. Cristofaro, “Development of a wearable device for sign language translation,” in Human friendly robotics, pp. 115–126, Springer, 2019.

[13] Q. Zhang, D. Wang, R. Zhao, and Y. Yu, “Myosign: enabling end-to-end sign language recognition with wearables,” in Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 650–660, 2019.

[14] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, “Wearable sensor-based sign language recognition: A comprehensive review,” IEEE Reviews in Biomedical Engineering, 2020.

[15] S. G. Azar and H. Seyedarabi, “Trajectory-based recognition of dynamic persian sign language using hidden markov model,” Computer Speech & Language, vol. 61, p. 101053, 2020.

[16] M. Ahmed, B. Zaidan, A. Zaidan, M. M. Salih, Z. Al-qaysi, and A. Alamoodi, “Based on wearable sensory device in 3d-printed humanoid: A new real-time sign language recognition system,” Measurement, vol. 168, p. 108431, 2021.

[17] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, “Signfi: Sign language recognition using wifi,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 1, pp. 1–21, 2018.

[18] H. Abdelnasser, K. Harras, and M. Youssf, “A ubiquitous wifi-based fine-grained gesture recognition system,” IEEE Transactions on Mobile Computing, vol. 18, no. 11, pp. 2474–2487, 2018.

[19] L. Zhang, Y. Zhang, and X. Zheng, “Wisign: Ubiquitous american sign language recognition using commercial wi-fi devices,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 11, no. 3, pp. 1–24, 2020.

[20] R. Cui, H. Liu, and C. Zhang, “A deep neural framework for continuous sign language recognition by iterative training,” IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1880–1891, 2019.

[21] E. K. Kumar, P. Kishore, M. T. K. Kumar, and D. A. Kumar, “3d sign language recognition with joint distance and angular coded color topographical descriptor on a 2-stream cnn,” Neurocomputing, vol. 372, pp. 40–54, 2020.

[22] R. Rastgou, K. Kiani, and S. Escalera, “Hand pose aware multimodal isolated sign language recognition,” Multimedia Tools and Applications, vol. 77, no. 37, 2018.

[23] N. B. Ibrahim, M. M. Selim, and H. H. Zayed, “An automatic arabic sign language recognition system (arsls),” Journal of King Saud University-Computer and Information Sciences, vol. 30, no. 4, pp. 470–477, 2018.

[24] Q. Xue, X. Li, D. Wang, and W. Zhang, “Deep forest-based monocular visual sign language recognition,” Applied Sciences, vol. 9, no. 9, p. 1945, 2019.

[25] H. Kolivand, S. Joudaki, M. S. Sunar, and D. Tully, “A new framework for sign language alphabet hand posture recognition using geometrical features through artificial neural network (part 1),” Neural Computing and Applications, pp. 1–19, 2020.

[26] K. M. Lim, A. W. C. Tan, C. P. Lee, and S. C. Tan, “Isolated sign language recognition using convolutional neural network hand modelling and hand energy image,” Multimedia Tools and Applications, vol. 78, no. 14, pp. 19917–19944, 2019.

[27] J. J. Tran, J. Kim, J. Chon, E. A. Riskin, R. E. Ladner, and J. O. Wobbrock, “Evaluating quality and comprehension of real-time sign language video on mobile phones,” in The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility, pp. 115–122, 2011.

[28] J. J. Tran, E. A. Riskin, R. E. Ladner, and J. O. Wobbrock, “Evaluating intelligibility and battery drain of mobile sign language video transmitted at low frame rates and bit rates,” ACM Transactions on Accessible Computing (TACCESS), vol. 7, no. 3, pp. 1–26, 2015.

[29] T. Sharma, S. Kumar, N. Yadav, K. Sharma, and P. Bhardwaj, “Air-swipe gesture recognition using opencv in android devices,” in 2017 international conference on algorithms, methodology, models and applications in emerging technologies (ICAMMAET), pp. 1–6, IEEE, 2017.

[30] G. A. Rao and P. Kishore, “Selfie video based continuous indian sign language recognition system,” Ain Shams Engineering Journal, vol. 9, no. 4, pp. 1929–1939, 2018.

[31] R. Jitcharoenpor, P. Senechakr, M. Dahan, A. Suchato, E. Chuangswanich, and P. Punyabukkana, “Recognizing words in thai sign language using flex sensors and gyroscopes,” in I-CREATE2017, vol. 4, 2017.

[32] M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, and M. M. h. Lakulu, “A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017,” Sensors, vol. 18, no. 7, pp. 2208, 2018.

[33] Y.-C. Chu, Y.-J. Jiang, T.-M. Tai, and W.-J. Hwang, “Recognition of hand gesture sequences by accelerometers and gyroscopes,” Applied Sciences, vol. 10, no. 18, pp. 5650, 2020.

[34] S. Jiang, L. Li, H. Xu, J. Xu, G. Gu, and B. P. Shull, “Stretchedable e-skin patch for gesture recognition on the back of the hand,” IEEE Transactions on Industrial Electronics, vol. 67, no. 1, pp. 647–657, 2019.

[35] A. Vaitkevičius, M. Taroza, T. Blažauskas, R. Damaševičius, R. Maskeliūnas, and M. Woźniak, “Recognition of american sign language gestures in a virtual reality using leap motion,” Applied Sciences, vol. 9, no. 3, pp. 445, 2019.

[36] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra, “A position and rotation invariant framework for sign language recognition (slr) using kinect,” Multimedia Tools and Applications, vol. 77, no. 7, pp. 8823–8846, 2018.

[37] V. Belissen, A. Braffort, and M. Gouiffes, “Dicta-sign-lsf-v2: remake of a continuous french sign language dialogue corpus and a first baseline for...
automatic sign language processing,” in LREC 2020, 12th Conference on Language Resources and Evaluation, 2020.

[38] I. Kagirol, D. Ivanov, D. Ryumin, A. Askoyev, and A. Karpov, “Therusian: Database of Russian sign language and motion,” in Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6079–6085, 2020.

[39] M. Kawulok and J. Nalepa, “Hand pose estimation using support vector machines with evolutionary training,” in IWSSIP 2014 Proceedings, pp. 87–90, IEEE, 2014.

[40] M. A. Almasre and H. Al-Nuaim, “A comparison of arabic sign language dynamic gesture recognition models,” Heliyon, vol. 6, no. 3, p. e03554, 2020.

[41] F. Zhang, S. Han, H. Gao, and T. Wang, “A gaussian mixture based hidden markov model for motion recognition with 3d vision device,” Computers & Electrical Engineering, vol. 83, p. 106603, 2020.

[42] S. Tornay, O. Aaran, and M. M. Dor, “An hmm approach with inherent model selection for sign language and gesture recognition,” in Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6049–6056, 2020.

[43] E. P. da Silva, P. D. P. Costa, K. M. O. Kamada, J. M. De Martino, and G. A. Florenzino, “Recognition of affective and grammatical facial expressions: a study for brazilian sign language,” in European Conference on Computer Vision, pp. 218–236, Springer, 2020.

[44] K. Polat and M. Saraçlar, “Unsupervised discovery of sign terms by k-nearest neighbours approach,” in European Conference on Computer Vision, pp. 310–321, Springer, 2020.

[45] M. Parello, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, “Exploiting 3d hand pose estimation in deep learning-based sign language recognition from rgb videos,” in European Conference on Computer Vision, pp. 249–263, Springer, 2020.

[46] N. C. Tamer and M. Saraçlar, “Improving keyword search performance in sign language with hand shape features,” in European Conference on Computer Vision, pp. 322–333, Springer, 2020.

[47] M. De Coster, M. Van Herreweghe, and D. Damon, “Sign language recognition with transformer networks,” in 12th International Conference on Language Resources and Evaluation, pp. 6018–6024, European Language Resources Association (ELRA), 2020.

[48] Q. Zhang, Y. Zhang, and Z. Liu, “A dynamic hand gesture recognition algorithm based on csi and yolov3,” in Journal of Physics: Conference Series, vol. 1267, p. 012055, IOP Publishing, 2019.

[49] S. Yuan, M. Han, L. Zhang, J. Lv, and F. Zhang, “Research approach of hand gesture recognition based on improved yolov3 network and bayes classifier,” in 2020 The 4th International Conference on Video and Image Processing, pp. 140–146, 2020.

[50] A. Mujahid, M. J. Awan, A. Yasin, M. A. Mohammed, R. Damačević, R. Maskeliūnas, and K. H. Aulkarkeem, “Real-time hand gesture recognition based on deep learning yolov3 model,” Applied Sciences, vol. 11, no. 9, p. 4164, 2021.

[51] 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.

[52] H. Bull, M. Gouiffès, and A. Briffa, “Automatic segmentation of sign language into subtitle-units,” in European Conference on Computer Vision, pp. 186–198, Springer, 2020.

[53] M. Borg and K. P. Camilleri, “Phonologically-meaningful subunits for deep learning-based sign language recognition,” in European Conference on Computer Vision, pp. 199–217, Springer, 2020.

[54] H. Bull, A. Braffort, and M. Gouiffès, “Mediapik-skela 2d skeleton video database of french sign language with aligned french subtitles,” in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 6063–6068, 2020.

[55] M. Kazmaarek and M. Filhol, “Alignment data base for a sign language concordancer,” in Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6069–6072, 2020.

[56] M. Mukushev, A. Saburov, A. Imashev, K. Koishibay, V. Kimmelman, and A. Sandygulova, “Evaluation of manual and non-manual components for sign language recognition,” in Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association (ELRA), 2020.

[57] F. Zhang, S. Han, H. Gao, and T. Wang, “A gaussian mixture based hidden markov model for motion recognition with 3d vision device,” Computers & Electrical Engineering, vol. 83, p. 106603, 2020.

[58] S. Paris, “Fast gmm and fisher vectors,” MATLAB Central File Exchange, 2021.

[59] D. Avoila, L. Cinque, M. De Marsico, A. Fagioli, and G. L. Foresti, “Lietome: Preliminary study on hand gestures for deception detection via fisher-lstm,” Pattern Recognition Letters, vol. 138, pp. 455–461, 2020.

[60] G. Keren and B. Schuller, “Convolutional rnn: an enhanced model for extracting features from sequential data,” in 2016 International Joint Conference on Neural Networks (IJCNN), pp. 3412–3419, IEEE, 2016.

[61] C. K. Lee, K. K. Ng, C.-H. Chen, H. C. Lau, S. Chung, and T. Tsoi, “American sign language recognition and training method with recurrent neural network,” Expert Systems with Applications, vol. 167, p. 114403, 2021.

[62] R. Rastgoo, K. Kiani, and S. Escalera, “Real-time isolated hand sign language recognition using deep networks and sgd,” Journal of Ambient Intelligence and Humanized Computing, pp. 1–21, 2021.

[63] S. Y. Boulahia, E. Anquetil, F. Multon, and R. Kulpa, “Dynamic hand gesture recognition based on 3d pattern assembled trajectories,” in 2017 seventh international conference on image processing theory, tools and applications (IPTA), pp. 1–6, IEEE, 2017.

[64] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, “Shrec’17 track: 3d hand gesture recognition using a depth and skeletal dataset,” in 3DOR-10th Eurographics Workshop on 3D Object Retrieval, pp. 1–6, 2017.

[65] R. A. Waltz, J. L. Morales, J. Nocoled, and D. Orban, “An interior algorithm for nonlinear optimization that combines line search and trust region steps,” Mathematical programming, vol. 107, no. 3, pp. 391–408, 2006.

[66] J. Nocoled and S. Wright, Numerical optimization. Springer Science & Business Media, 2006.

[67] J. Ngiam, Z. Chen, S. Bhaskar, P. Koh, and A. Ng, “Sparse filtering,” Advances in neural information processing systems, vol. 24, pp. 1125–1133, 2011.

[68] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, “Skeleton-based dynamic hand gesture recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9, 2016.

[69] E. Ohn-Bar and M. Trivedi, “Joint angles similarities and hog2 for action recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 465–470, 2013.

[70] O. Oreifej and Z. Liu, “Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 716–723, 2013.

[71] M. Devanne, H. Wannous, S. Berrett, P. Pala, M. Daoudi, and A. Del Bimbo, “3-d human action recognition by shape analysis of motion trajectories on riemannian manifold,” IEEE transactions on cybernetics, vol. 45, no. 7, pp. 1340–1352, 2014.

[72] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, “Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition,” in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0, 2019.

[73] J. Liu, Y. Liu, Y. Wang, V. Prinet, S. Xiang, and C. Pan, “Decoupled representation learning for skeleton-based gesture recognition,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5751–5760, 2020.

[74] K. Lupinet, A. Ranieri, F. Giannini, and M. Monti, “3d dynamic hand gestures recognition using the leap motion sensor and convolutional neural networks,” in International Conference on Augmented Reality, Virtual Reality and Computer Graphics, pp. 420–439, Springer, 2020.

[75] B. Hisham and A. Hamouda, “Arabic sign language recognition using ada-boosting based on a leap motion controller,” International Journal of Information Technology, vol. 13, no. 3, pp. 1221–1234, 2021.