IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 30, 2021, accepted January 27, 2022, date of publication February 2, 2022, date of current version February 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3148132

American Sign Language Words Recognition Using Spatio-Temporal Prosodic and Angle Features: A Sequential Learning Approach

SUNUSI BALA ABDULLAHI[1,2], (Member, IEEE),
AND KOSIN CHAMNONGTHAI[3], (Senior Member, IEEE)
[1]Department of Computer Engineering, Faculty of Engineering, King Mongkut's Univeristy of Technology Thonburi, Bangkok 10140, Thailand
[2]Force Criminal Investigation and Intelligence Department, Nigeria Police, Maitama, Abuja 900211, Nigeria
[3]Department of Electronic and Telecommunication Engineering, Faculty of Engineering, King Mongkut's Univeristy of Technology Thonburi, Bangkok 10140, Thailand

Corresponding author: Kosin Chamnongthai (kosin.cha@kmutt.ac.th)

This work was supported by the Petchra Pra Jom Klao Ph.D. Research Scholarship from the King Mongkut's University of Technology Thonburi, Bangkok, Thailand, under Grant 45/2562.

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 an 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 98.6% 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. Hard of hearing are generally dependent on sign language to participate in the real world. World Federation of the hard of hearing 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 hard of hearing to express themselves. Therefore, these words are indispensable for daily hard of hearing communication. It is imperative to

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

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 μ 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, FV 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 joints 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**), **(c)**, **(f)** and 3(**a**) and **(d)**. However, small motion at wrist generate large variation angles (Δ_φ). 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 feature, 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 between finger joints and fingertips ($Z_{j,k,l}$) as refers in [1], equivalently written as

$$Z_{j,k,l} = [t's/po(j3), tj/tk(j3, tj), t's/j's(t, j)]. \quad (1)$$

TABLE 1. SLR according to capturing modalities.

Algorithm name	Brief methodology	Highlights	Limitations
		RF-based SLR methods	
Talking Hands [12]	Distance function + glove + smartphone speech synthesizer	RF via scenario translation	Intrusive, bulk and unnatural fails to dynamic words
MyoSign [13]	Myo arm band signal Multi-CNN + BiLSTM + CTC	Real-time No temporal segmentation	unnatural occlusion
Seyedarabi et. Al. [15]	White glove signal HMM + Gaussian	hand shape + trajectory region growing technique	cumbersome occlusion + intrusive
Data glove [16]	Data glove + IMU + TOF + FSR	Real-time + 3D printed humanoid Hand kinematics + SL analysis	cumbersome + tuning is time + consuming + invasive
WiSign [19]	WiFi signal + CSI + PSD DBN + HMM-Gaussian	language model Multimodal gestures	high internet access non-ubiquitous
SignFi [17]	WiFi signal + CSI WiFi packets + CNN	CSI measurements Multimodal gestures	internet access + interference non-ubiquitous
WiGest [18]	WiFi signal + CSI DWT + Gaussian noise-SURE	No gesture learning Ubiquious system	high internet access interference
		Digital camera and camera stereo SLR methods	
ArSLRS [23]	RGB videos + YCbCr color space Euclidean distance	YCbCr segmentation multimodal fusion	segmentation complexity complex environment + skin effect
Xue et al. [24]	Multimodal RGB + OpenPose Voting strategy + deep forest	skeleton projection Semantic consideration	skin effect complex learning
JDTD and JATD [20]	Multimodal RGB + OF images 3D motion camera information	two stream of CNN	cumbersome + invasive skin effect
DNN [21]	3D motion camera information RGB	End-to-end learning multimodal fusion	skin effect cumbersome
Rastgoo et al. [22]	Multimodal RGB + Depth LSTM + CNN	3D multimodal fusion optical + scene flow	segmentation complexity 2D projection looks alike
ASLNN [25]	Depth sensor camera + DGSLR	LSM + CHA + CHP + ANN	hard pre- and post processing
Tran et al. [27], [28]	Smartphone-based capturing	Human Signal intelligibility model	static gestures + hard processing
Air-Swipe Gesture [29]	Smartphone-based capturing	Ubiquitous SLR + OpenCV	compression + segmentation
Selfie SLRs [30]	Smartphone-based + DCT + PCA MDC + Euclidean distance	Selfie-based capturing Gaussian pre-filtering	preprocessing complexity Not applicable while walking
Lim et al. [26]	particle filter + HEI + GEI + CNN	hybrid hand modeling	segmentation complexity
Dicta-Sign-LSF-v2 [37]	3 cameras + CRNN	SLP from Iconic structure DB of 35000 manual units 720 x 575 at 25 fps	multimodality
		Sensor-based SLR methods	
Jitcharoenport [32]	fLT + LDA + k-NN	sensor-based capturing	calibration + cumbersome low accuracy
Chu et al. [33]	Residual PairNets + MAP	Accelerometers + gyroscopes	cumbersome
Stretchable e-skin [34]	Backhand-view based capturing mutual information + LDA	motion camera multiple sensors	pervasive + trial and error unnatural
		Active imaging device SLR methods	
Kumar et al. [36]	Kinect skeleton coordinate + HMM	Real-time position invariant system	hard learning + Limited FoV
Liu and Huai [1]	LMC + HMM-PSO	Dynamic hand gestures	hard learning
Aurelijus et al. [35]	LMC + HMC	microservice via internet recognition	limited representation ability
DeepASl [5]	Backhand-view based LMC + HBRNN	Ubiquitous + Real-time SLR	misclassification
TheRuSLan [38]	MS Kinect 2.0 + Kinect SDK	DB of 3D RSL lexical units 1920 x 1080 pixels at 30 fps	SL database design

TABLE 2. Sign language recognition methods.

Algorithm name	Brief methodology	Highlights	Limitations
Kawulok and Nalepa [39]	SVM + evolutionary strategy	vector differences between shapes	separate features learning
Almasre and Al-Nuaim [40]	DPM with SVM + RF + KNN	dynamic gesture recognition	low representation ability
GMM-HMM [41]	WLR + GMM-HMM	key frames denotes hidden states	computationally hard
Tornay et al. [42]	continuous + HMM	each sign denotes CHMM states	separate features extraction
da Silva et al. [43]	CNN + CNN-LSTM	joint region recognition of AU Video analysis of AUs with FACS	manual features not fully preserve learning small variation fails
Polat and Saraclar [44]	UTD + KNN	RGB videos from OP and HD signer dependent approach	Segmentation complexity limited learning ability
Parelli et al. [45]	Attention-based CNN	3D skeletal information OpenPose from RGB videos	information may looks alike decrease in accuracy due to ED
KWS [46]	KWS + end-to-end CNN	Relationship between signs and spoken words. Mixing pose and shape KWS of OP	Segmentation constrained condition of fusion model
De Coster et al. [47]	OpenPose + MTNs	SL annotation	misclassification
Zhang et al. [48]	WF + KF + CSI + YOLOv3	Continuous dynamic gestures	preprocessing complexity
Yuan et al. [49]	YOLOv3-STN + PCA-Bayes	improved YOLOv3 and Bayes	computationally hard
Mujahid et al. [50]	YOLOv3 + DarkNet-53	end-to-end learning of static of static gestures. Real-time	static gesture recognition
LSTM2+CHMM [51]	LSTM2 + CHMM + CNN + graph	hand segmentation using PRT RGB-D image fusion	Segmentation complexity
Avola et al. [6]	Multi-stack LSTM	angle and displacement of skeletal finger joints	misclassification
Bull et al. [52]	ST-GCN + BiLSTM	detection of temporal boundaries of subtitles	segmentation issues
Borg et al. [53]	OpenPose + factorization NRSfM + Bi-RNN-CTC	subunits model-based SLR preserve phonology meaning employ trajectory features	Bi-RNN sometimes may lead to memory explosion especially when the sequence grows
MEDIAPI-SKEL 2D [54]	OpenPose + concodancer + GCN	skeleton-based approach Integrate CAT in SL	Study framework
Kaczmarek and Filhol [55]	Brat + Elan + CAT	SL annotation + alignment	SL alignment framework
Mukushev et al. [56]	OpenPose + Logistic regression	non-manual continuous SL none deep learning method	LOG models are good at monotonic relationships

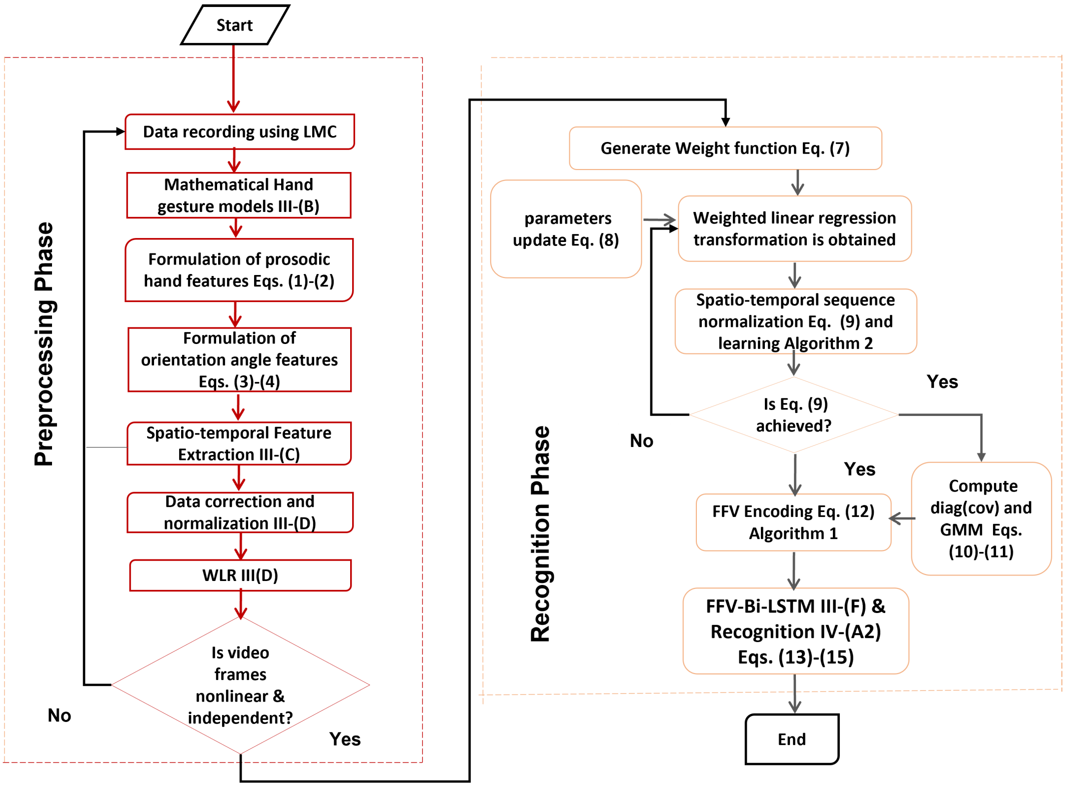
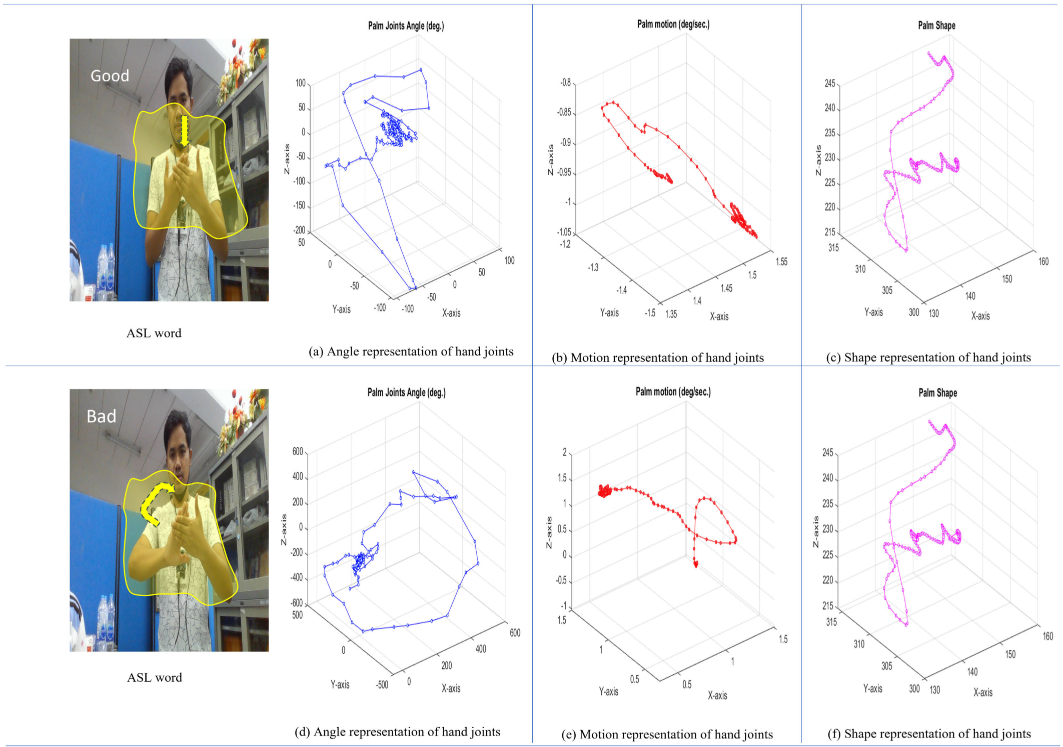

Start
Preprocessing Phase
Data recording using LMC
Mathematical Hand gesture models III-(B)
Formulation of prosodic hand features Eqs. (1)-(2)
Formulation of orientation angle features Eqs. (3)-(4)
Spatio-temporal Feature Extraction III-(C)
Data correction and normalization III-(D)
WLR III(D)
Is video frames nonlinear & independent?
No
Yes
Recognition Phase
Generate Weight function Eq. (7)
parameters update Eq. (8)
Weighted linear regression transformation is obtained
Spatio-temporal sequence normalization Eq. (9) and learning Algorithm 2
Is Eq. (9) achieved?
No
Yes
Yes
Compute diag(cov) and GMM Eqs. (10)-(11)
FFV Encoding Eq. (12) Algorithm 1
FFV-Bi-LSTM III-(F) & Recognition IV-(A2) Eqs. (13)-(15)
End

FIGURE 1. Flow chart of the proposed method.



(a) Angle representation of hand joints

(b) Motion representation of hand joints

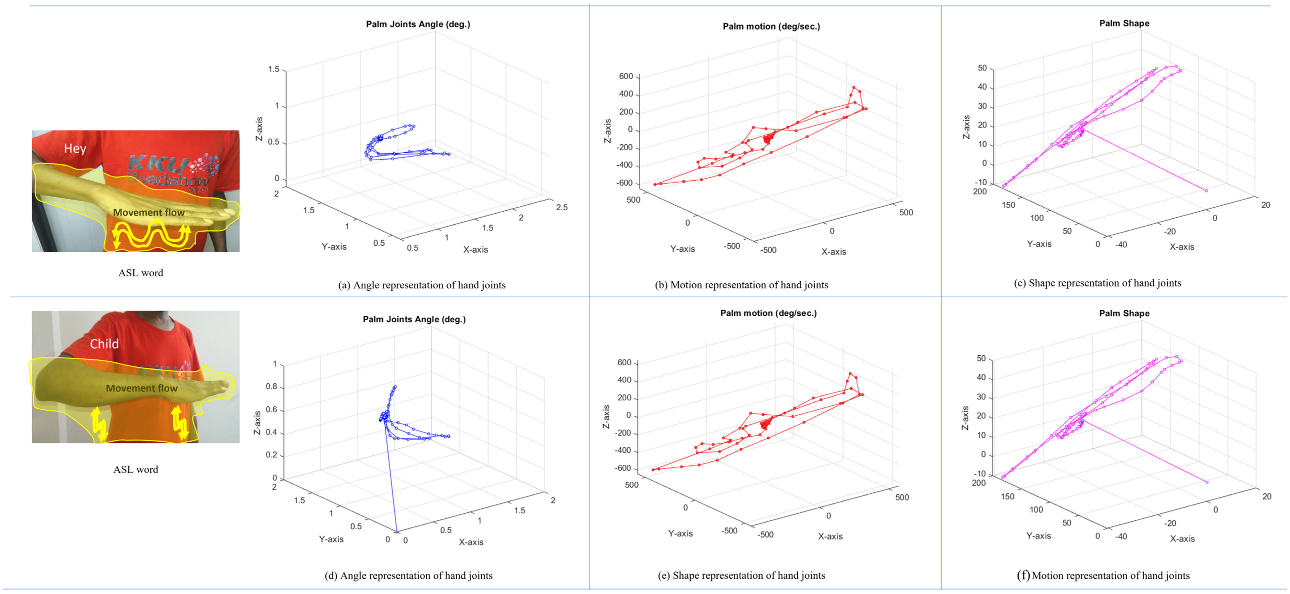
(c) Shape representation of hand joints

(d) Angle representation of hand joints

(e) Motion representation of hand joints

(f) Shape representation of hand joints

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.



(a) Angle representation of hand joints

(b) Motion representation of hand joints

(c) Shape representation of hand joints

(d) Angle representation of hand joints

(e) Shape representation of hand joints

(f) Motion representation of hand joints

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.

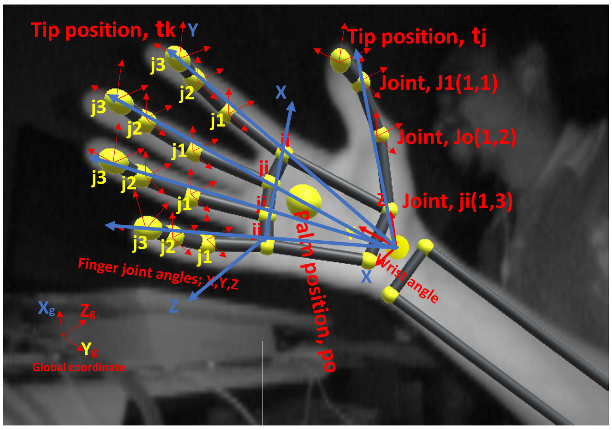
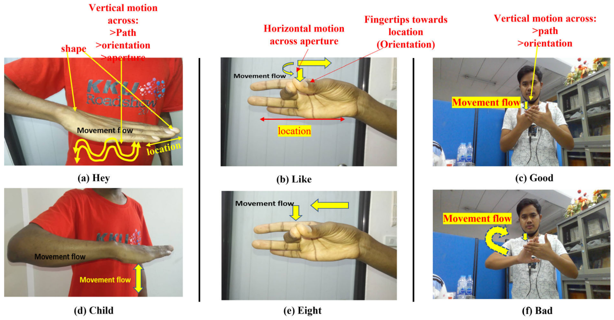


FIGURE 4. Skeleton hand joints definitions.


Vertical motion across: >Path >orientation >aperture
shape
Movement flow
location
Horizontal motion across aperture
Fingertips towards location (Orientation)
Movement flow
location
Vertical motion across: >path >orientation
Movement flow
Movement flow
Movement flow
Movement flow
Movement flow

(a) Hey

(b) Like

(c) Good

(d) Child

(e) Eight

(f) Bad

FIGURE 5. Similar ASL words using single hand according to prosodic model.

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 μ of finger joints motion $M^f(n)$ per each frame f can be coined as $\upsilon^f(n)$, where n denotes number of sequence per each frame. Thus,

$$\mu = \{M^f + \upsilon^f + Z^f\}. \tag{2}$$

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

$$Y_R = \frac{Z_R \times X_L}{|Z_R \times X_L|}. \tag{3}$$

Thus, angle between Z_R and X_L, is denoted as a. Wrist flexion and extension angle is denoted as φ. Similarly, hand internal angles b [6] 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

$$Q = \{\varphi + a + b\}. \tag{4}$$

Therefore, extracted features according to formulations in Eqs. (1)-(4) are fused through simple vector concatenation equivalently written as:

$$\lambda_v = [\mu^f(n), Q^f(n), \rho] \tag{5}$$

where $v \in [thumb, index, middle, ring, little]$, and ρ 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

$$L = [M_1, M_2, \cdots, M_F] \tag{6}$$

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). This value is assumed from hand motion velocity (which is $\geq 45\%$). 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 $\geq 45\%$ of maximum velocity (peak velocity), as illustrated in frames **(b)** and **(e)** of Figs. 2 - 3. Hand velocity, orientation angles, spatial and inherent features are computed per 22 hand joints together to make a sum of 5×32 (192) information per frame.

D. DATA CORRECTION AND NORMALIZATION

The output obtained from setup illustrated in Fig. 10 contains noise, which is handled by Savitzky-Golay smoothing filter. The smoothed information $B_{h,s,f} = \{b^k_{p,q,r,f}, \cdots, b^k_{p,q,r,n}\}$ is utilized using local weighted linear regression (WLR) algorithm to handle missing values and nonlinearity [57]. Thus, weight function is added into linear regression as follow

$$\eta(w_f) = e^{-}\frac{(w - w_f)}{2\lambda^2} \tag{7}$$

where w denotes prediction time, w_f denotes data progressing time and λ denotes wavelength parameter. Then, parameter update is given in Eq. (8) and results of corrected video information is illustrated in Fig. 6.

$$\theta^d_k = \theta^{d+1}_k + \alpha \sum_{f=1}^{n} \eta(w_f)(b^k_f - O_\theta(w_f))w^d_f. \tag{8}$$

Furthermore, after data correction *Y(w)*, then there is need to normalize hands by zero centering wrist, using the following equation

$$\beta_{w,f} = \begin{cases} (0,0,0), & if \quad f = 1. \\ Y_{w,f} - Y_{w=Right,f-1}, & if \quad f = 2, \cdots, F. \end{cases} \tag{9}$$

E. FAST FISHER VECTOR ENCODING (FFV)

FFV transform the features by their deflection from a generative model (Gaussian Mixture Model (GMM)) using sparse matrix representation (sparse filtering [58]). GMM is utilized as probability density function with mixture weight (w), mean vector (μ), parameters θ, and covariance matrix (*diag(cov)*) 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:

$\theta = \{w_k, \mu_k, \sum_{k=1}^{K} = diag(cov_1 k, \cdots, cov_t k) : k = 1 \cdots K\}$. To apply FFV ($\varkappa$) to our features, let $\lambda = \{\lambda_t : t = 1 \cdots T\}$ be the set of T local information in Eq. (8), thus, generative procedures λ of the whole feature vectors are formulated as follows

$$H_\theta(\lambda) = \frac{1}{v}\sum_{k=1}^{K}(\lambda_t; \mu_k, cov_k)w_k, \tag{10}$$

Algorithm 1 Fisher Vector Transformation

```
 1: start
 2: set $\beta$ {Video features}
 3: set $\mu, \sigma, v, k, diag(cov.)$ {GMM parameters}
 4: set $\psi$ {Target features}
 5: repeat
 6:     while $\beta$ is detect do
 7:     normalize $\beta$
 8:     set $j = 0$ to sequence length
 9:     set $l = 0$ to sequence length
10:     end while
11:     compute $\mu, \sigma, v, k, diag(cov.)$ via EM
12:     compute $f = \beta(features[j], \mu_l, cov._l)$
13:     compute G from Eq. (11)
14:     while H best fit $\beta$ from Eq. (10) do
15:     Get FFV encoding using step 11
16:     end while
17:     compute Eq. (12)
18: until Eq. (11) converge
19: return Eq. (11)
20: end
```

Also, FFV matrix can be obtained as follows:

$$\varkappa_\lambda = [\nabla_\theta log\mu_\theta(T)\nabla_\theta log\mu_\theta\overline{T}]. \tag{11}$$

Similarly, $\varkappa$ is finally obtained from fused partial derivatives through GMM parameters

$$\varkappa^t = [G^t_{\mu,1}, G^t_{s,1}, \cdots G^t_{\mu,k} G^t_{s,k}]. \tag{12}$$

where H_θ, $1/v$, $\nabla_\theta log(\cdot)$ denote generative model parameters, normalized values, and log-likelihood gradient. The θ are discover from training features via expectation maximization (EM) strategy. Gradients are computed according to mean vector μ_f and standard deviation (s_k) of the *fth* Gaussian in Eq. (12).

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 *L2* normalization throughout Gaussians and λ as follows

$$\begin{aligned} O_{f,\varkappa} = \sigma[& V_{\overrightarrow{h}o}\overrightarrow{h}_f, Q_{\varkappa f} + V_{\overleftarrow{h}o}\overleftarrow{h}_f, Q_{\varkappa f} + V_{\overrightarrow{h}o}\overrightarrow{h}_f, \mu_{\varkappa f} \\ & + V_{\overleftarrow{h}o}\overleftarrow{h}_f, \mu_{\varkappa f} + V_{\overrightarrow{h}o}\overrightarrow{h}_f, \rho_{\varkappa f} \\ & + V_{\overleftarrow{h}o}\overleftarrow{h}_f, \rho_{\varkappa f} + V_{\overrightarrow{h}o}\overrightarrow{h}_f, L_{\varkappa f} \\ & + V_{\overleftarrow{h}o}\overleftarrow{h}_f, L_{\varkappa f} + d_o] \end{aligned} \tag{13}$$

where σ, V_{ho}, h_f, Q, μ, ρ, and L denotes logistic sigmoid function, weight matrices, angle, motion, shape, and spatial

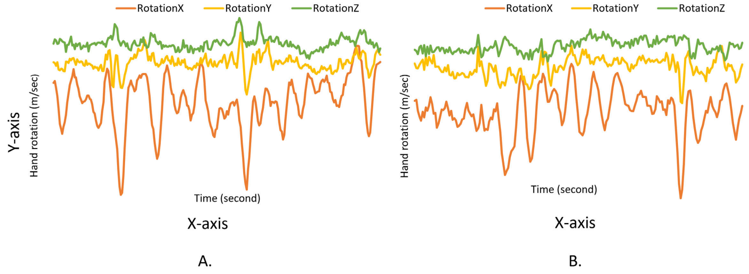


FIGURE 6. Data correction: A. shows original average skeletal hand video frames, and B. represents smoothed and corrected frames information using weighted linear regression.

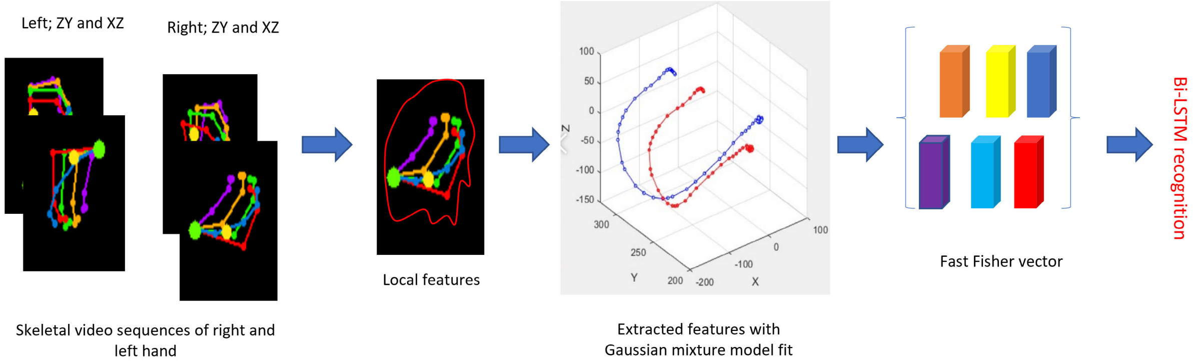

Left; ZY and XZ
Right; ZY and XZ
Skeletal video sequences of right and left hand
Local features
Extracted features with Gaussian mixture model fit
Fast Fisher vector
Bi-LSTM recognition
Z
Y
X

FIGURE 7. 3D keypoints generation with Fast Fisher vector transformation.

TABLE 3. Simulation environment.

Systems	Requirements
Personal Computer	Dell G3 15 Gaming CPU: Iintel 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	10 persons frequency: 10 times per word

TABLE 4. Network parameter settings.

Network layers	Parameter options	Selection
Input layer	Sequence length	Longest
	Batch size	27
	Input per sequence	168
	Feature vector	1 dimension
Hidden layer	Bi-LSTM layer	Longest
	Hidden state	40
	Activation function	Softmax
Output layer	LSTM model	Many to one
	Number of class	10

features encoded by FFV $\varkappa^f$, and d_o denotes bias. Where $\overrightarrow{h_o}$, and $\overrightarrow{h_f}$ denotes forward hidden and cell state vectors. $\overleftarrow{h_o}$ and $\overleftarrow{h_f}$ 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.

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 they perform the sign while walking and standing.

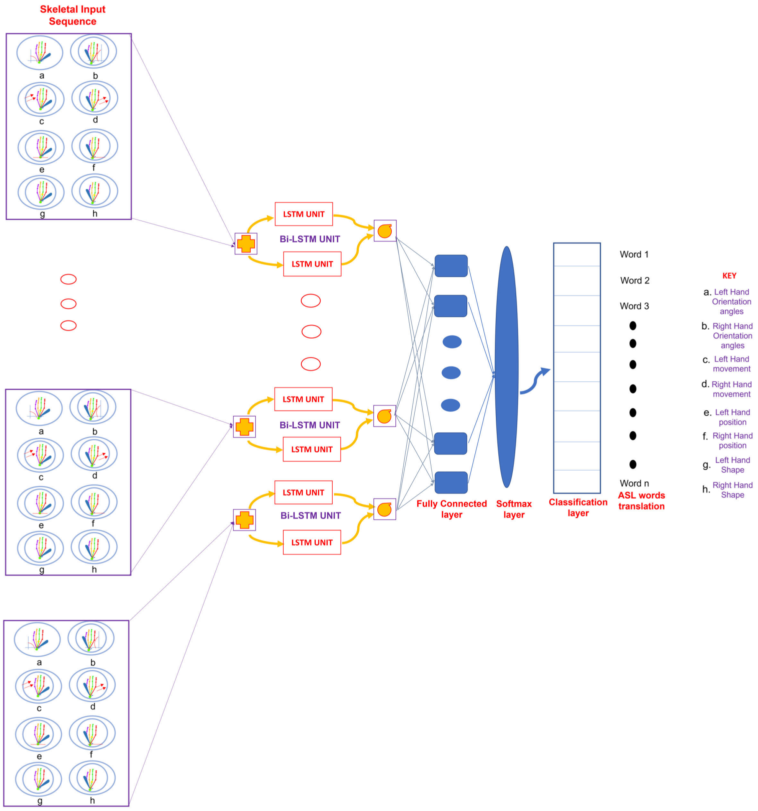


FIGURE 8. Architecture of Bi-directional LSTM.

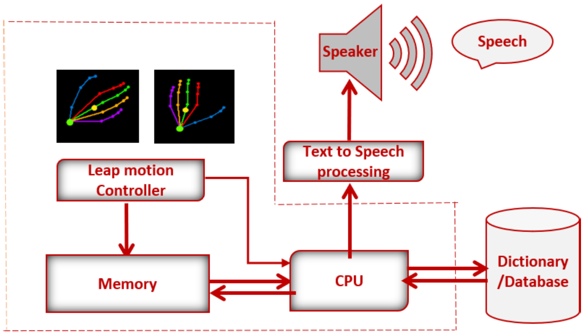


FIGURE 9. Block diagram of needed hardware components.

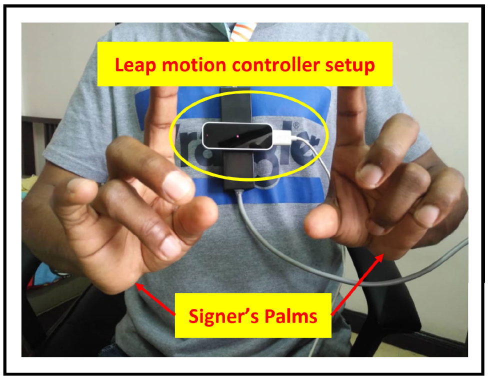


FIGURE 10. Photo of experimental system.

Each signer perform 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 daily used first 100 ASL words. Some example of our datasets are given in Fig. 5. The data set 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] Data set, respectively.

2) RECOGNITION PHASE

Our algorithm call a function *InitialTransformWeights* name-value pair. Sparse filtering algorithm is implemented in MATLAB using "*sparsefilt*" function from yael package. The algorithm handle 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 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

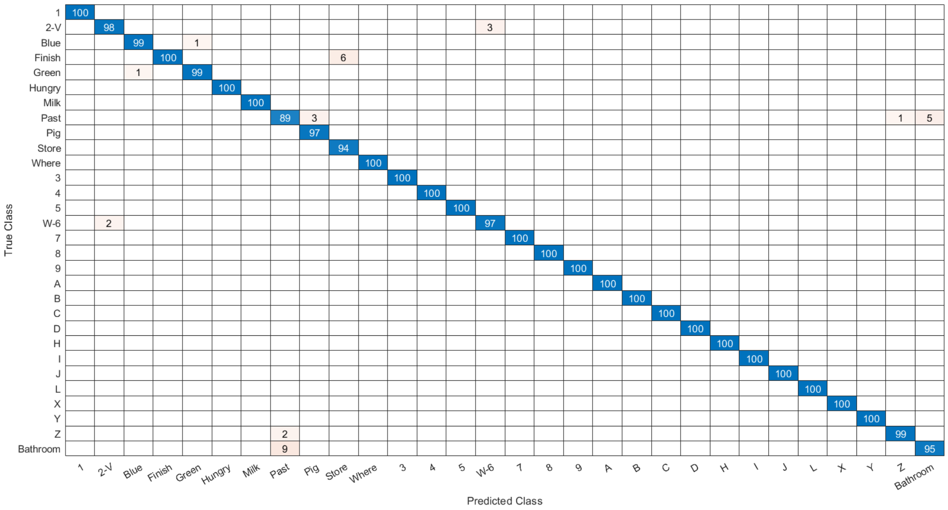


FIGURE 11. Confusion matrix of skeletal ASL datasets [6] using adopted method.

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

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_t^E 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:

$$O = \sum_{f=0}^{F-1} O_{ff}^{f} \tag{14}$$

$$O^L = p(E_l|G) = \frac{e^{O^l}}{\sum_{i=0}^{L-1} e^{O_i}}, \quad L = 1, \cdots, L \tag{15}$$

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

B. RESULTS

We reported performance results of FFV-Bi-LSTM algorithm. Overall comparison results between FFV-Bi-LSTM and Avola *et al.* [6] method are shown in Table 5. Average recognition of FFV-Bi-LSTM are illustrated in Table 9 for 10-pairs highly correlated ASL words and randomly selected 57 ASL words. The computational performance of FFV-Bi-LSTM in the proposed data set is depicted in Table 10. To show the effectiveness of the FFV optimization, we extend tests on spatio-temporal features without and with the FFV optimizations mentioned in subsection III-F, detailed as Tables 11 and 12 for "Bi-LSTM no FFV optimization" and "FFV-Bi-LSTM". It is, therefore, demonstrates that our adopted algorithm is feasible for ubiquitous applications. We compare the performance accuracy of FFV-Bi-LSTM

TABLE 6. Results comparison on SHREC dataset.

Approach	Accuracy (%)	
	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

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 words		91.002	10 pairs	9
DeepASL [5]	DeepASL	94.5	56	5.5

TABLE 9. Scores per recognized correlated ASL words.

S/no.	Class	Accuracy (%)	Error (%)
1	Child	90	10
2	Eight	100	0
3	Embarrassed	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	Cheap	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

with some existing state-of-the-art methods, the average recognition accuracy for each is plotted in Figs. 11 and 13 and the accuracy values (Precision, Recall and F-scores) are listed in Tables 8, 6, 5, 7.

Algorithm 2 Sequential Feature Learning

```
start
set $L$ in Eq. (6) {Video input sequence}
set $V_h$ {Sequence weight}
set $S$ {Sequence length}
set $n$ {Hand index}
for each $n \in [0, s-1]$ do
repeat
    if $n < s - 1$ then
    Feed $M_n and V_h$ to Bi-LSTM
    else if
    $n \leftarrow S - 1$ then
    Get $M_n$ from Eq. (6) {Features for Bi-LSTM}
    else if
    stop
    end if
    end for
    compute parameters and recognition metrics
until Eq. (14) converge
return Eq. (15)
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 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.

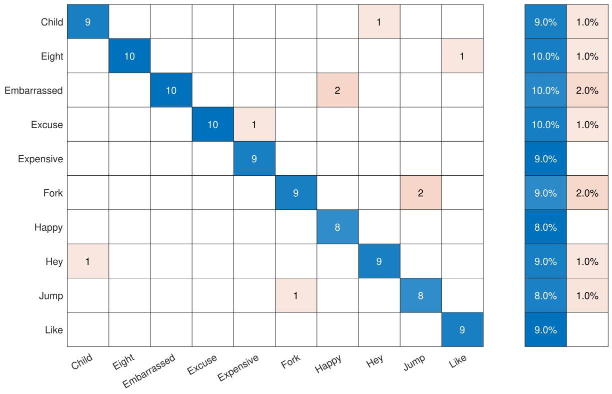


FIGURE 12. Confusion matrix of Correlated ASL words using adopted method.

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	9	10 pairs

TABLE 11. Different features combination for various Deep Bi-LSTM model comparison.

Epoch	minibatch size	Model combination	Iteration	Processing time (Train)	Processing time (Test)	Accuracy (%)	Learning rate
350	27	Shape + Motion	3500	371	85	76	1.00E-19
350	27	Shape + Motion + location	3500	753	199	81.38	1.00E-17
350	27	Shape + Motion + location + angular features	10500	5232	360	86.086	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

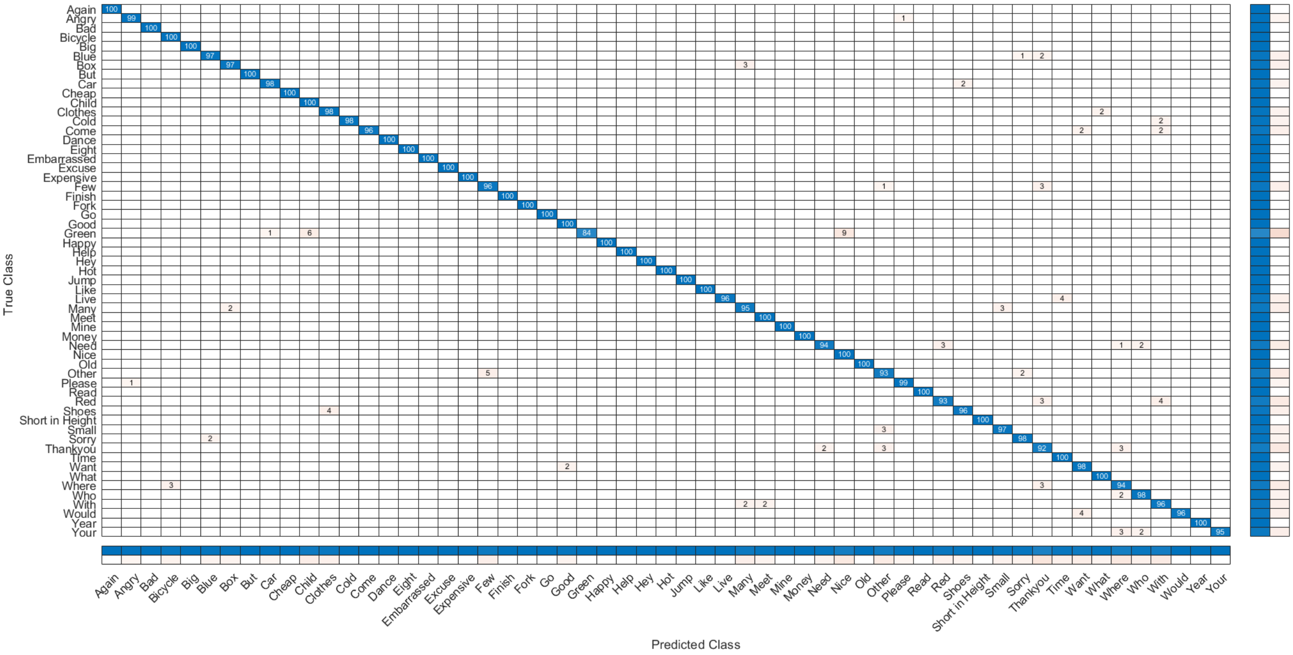


FIGURE 13. Confusion Matrix of the entire dataset.

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 our method with the method of [6]. FFV-Bi-LSTM algorithm fail to learn the small changes of hand motion trajectory of some similar ASL words, which reflect biases, which is responsible for misclassification. Since several features are influencing the recognition of similar ASL words, it is suggested that similar ASL words should be dealt with as a multi-feature problem in future research.

ACKNOWLEDGMENT

The authors 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," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 1, Jan. 2019, Art. no. 1950003.

[2] *With Sign Language, Everyone is Included*, World Fed. Deaf, Helsinki, Finland, 2018.

[3] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, Jan. 2019.

[4] P. Chophuk and K. Chamnongthai, "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 *Proc. 15th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2017, pp. 1–13.

[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 semaphoric hand gestures," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 234–245, Jan. 2019.

[7] S. Ameur, A. Ben Khalifa, and M. S. Bouhlel, "A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion," *Entertainment Comput.*, vol. 35, Aug. 2020, Art. no. 100373.

[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, no. 10, Oct. 2008, Art. no. P10008.

[9] R. Battison, *Lexical Borrowing in American Sign Language*. Silver Spring, MD, USA: Linstok Press, 1978.

[10] D. Brentari, *Sign Language Phonology*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[11] R. A. Tennant, M. Gluszak, and M. G. Brown, *The American Sign Language Handshape Dictionary*. Washington, DC, USA: Gallaudet Univ. Press, 1998.

[12] F. Pezzuoli, D. Corona, M. L. Corradini, and A. Cristofaro, "Development of a wearable device for sign language translation," in *Human Friendly Robotics*. Cham, Switzerland: Springer, 2019, pp. 115–126.

[13] Q. Zhang, D. Wang, R. Zhao, and Y. Yu, "MyoSign: Enabling end-to-end sign language recognition with wearables," in *Proc. 24th Int. Conf. Intell. User Interfaces*, Mar. 2019, pp. 650–660.

[14] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, "Wearable sensor-based sign language recognition: A comprehensive review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 82–97, 2021.

[15] S. G. Azar and H. Seyedarabi, "Trajectory-based recognition of dynamic Persian sign language using hidden Markov model," *Comput. Speech Lang.*, vol. 61, May 2020, Art. no. 101053.

[16] M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, Z. T. Al-Qaysi, and A. H. Alamoodi, "Based on wearable sensory device in 3D-printed humanoid: A new real-time sign language recognition system," *Measurement*, vol. 168, Jan. 2021, Art. no. 108431.

[17] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign language recognition using WiFi," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–21, Mar. 2018.

[18] H. Abdelnasser, K. Harras, and M. Youssef, "A ubiquitous WiFi-based fine-grained gesture recognition system," *IEEE Trans. Mobile Comput.*, vol. 18, no. 11, pp. 2474–2487, Nov. 2019.

[19] L. Zhang, Y. Zhang, and X. Zheng, "WiSign: Ubiquitous American sign language recognition using commercial Wi-Fi devices," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–24, Jun. 2020.

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

[21] E. K. Kumar, P. V. V. 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, Jan. 2020.

[22] R. Rastgoo, K. Kiani, and S. Escalera, "Hand pose aware multimodal isolated sign language recognition," *Multimedia Tools Appl.*, vol. 80, pp. 1–37, Sep. 2020.

[23] N. B. Ibrahim, M. M. Selim, and H. H. Zayed, "An automatic Arabic sign language recognition system (ArSLRS)," *J. King Saud Univ.-Comput. Inf. Sci.*, 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," *Appl. Sci.*, vol. 9, no. 9, p. 1945, May 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 Comput. Appl.*, vol. 33, pp. 1–19, May 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 Appl.*, vol. 78, no. 14, pp. 19917–19944, Jul. 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 *Proc. 13th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS)*, 2011, pp. 115–122.

[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 Trans. Accessible Comput.*, vol. 7, no. 3, pp. 1–26, Nov. 2015.

[29] T. Sharma, S. Kumar, N. Yadav, K. Sharma, and P. Bhardwaj, "Air-swipe gesture recognition using OpenCV in Android devices," in *Proc. Int. Conf. Algorithms, Methodology, Models Appl. Emerg. Technol. (ICAMMAET)*, Feb. 2017, pp. 1–6.

[30] G. A. Rao and P. V. V. Kishore, "Selfie video based continuous Indian sign language recognition system," *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 1929–1939, Dec. 2018.

[31] R. Jitcharoenpory, P. Senechakr, M. Dahlan, A. Suchato, E. Chuangsuwanich, and P. Punyabukkana, "Recognizing words in thai sign language using flex sensors and gyroscopes," in *Proc. i-CREATe*, vol. 4, 2017, pp. 1–4.

[32] M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, and M. M. B. 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, p. 2208, 2018.

[33] Y.-C. Chu, Y.-J. Jhang, T.-M. Tai, and W.-J. Hwang, "Recognition of hand gesture sequences by accelerometers and gyroscopes," *Appl. Sci.*, vol. 10, no. 18, p. 6507, Sep. 2020.

[34] S. Jiang, L. Li, H. Xu, J. Xu, G. Gu, and P. B. Shull, "Stretchable e-Skin patch for gesture recognition on the back of the hand," *IEEE Trans. Ind. Electron.*, vol. 67, no. 1, pp. 647–657, Jan. 2020.

[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," *Appl. Sci.*, vol. 9, no. 3, p. 445, Jan. 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 Appl.*, vol. 77, no. 7, pp. 8823–8846, Apr. 2018.

[37] V. Belissen, A. Braffort, and M. Gouiffès, "Dicta-Sign-LSF-v2: Remake of a continuous French sign language dialogue corpus and a first baseline for automatic sign language processing," in *Proc. 12th Conf. Lang. Resour. Eval. (LREC)*, 2020, pp. 1–12.

[38] I. Kagirov, D. Ivanko, D. Ryumin, A. Axyonov, and A. Karpov, "TheRuSLan: Database of Russian sign language," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 6079–6085.

[39] M. Kawulok and J. Nalepa, "Hand pose estimation using support vector machines with evolutionary training," in *Proc. IWSSIP*, 2014, pp. 87–90.

[40] M. A. Almasre and H. Al-Nuaim, "A comparison of Arabic sign language dynamic gesture recognition models," *Heliyon*, vol. 6, no. 3, Mar. 2020, Art. no. e03554.

[41] F. Zhang, S. Han, H. Gao, and T. Wang, "A Gaussian mixture based hidden Markov model for motion recognition with 3D vision device," *Comput. Electr. Eng.*, vol. 83, May 2020, Art. no. 106603.

[42] S. Tornay, O. Aran, and M. M. Doss, "An HMM approach with inherent model selection for sign language and gesture recognition," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 6049–6056.

[43] E. P. da Silva, P. D. P. Costa, K. M. O. Kumada, J. M. De Martino, and G. A. Florentino, "Recognition of affective and grammatical facial expressions: A study for Brazilian sign language," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 218–236.

[44] K. Polat and M. Saraçlar, "Unsupervised discovery of sign terms by k-nearest neighbours approach," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 310–321.

[45] M. Parelli, 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 *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 249–263.

[46] N. C. Tamer and M. Saraçlar, "Improving keyword search performance in sign language with hand shape features," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 322–333.
[47] M. De Coster, M. Van Herreweghe, and J. Dambre, "Sign language recognition with transformer networks," in *Proc. 12th Int. Conf. Lang. Resour. Eval.*, 2020, pp. 6018–6024.
[48] Q. Zhang, Y. Zhang, and Z. Liu, "A dynamic hand gesture recognition algorithm based on CSI and YOLOv3," *J. Phys., Conf. Ser.*, vol. 1267, Jul. 2019, Art. no. 012055.
[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 *Proc. 4th Int. Conf. Video Image Process.*, Dec. 2020, pp. 140–146.
[50] A. Mujahid, M. J. Awan, A. Yasin, M. A. Mohammed, R. Damaševičius, R. Maskeliūnas, and K. H. Abdulkareem, "Real-time hand gesture recognition based on deep learning YOLOv3 model," *Appl. Sci.*, vol. 11, no. 9, p. 4164, May 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. Braffort, "Automatic segmentation of sign language into subtitle-units," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 186–198.
[53] M. Borg and K. P. Camilleri, "Phonologically-meaningful subunits for deep learning-based sign language recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 199–217.
[54] H. Bull, A. Braffort, and M. Gouiffès, "MEDIAPI-SKEL—A 2D-skeleton video database of French sign language with aligned French subtitles," in *Proc. 12th Conf. Lang. Resour. Eval. (LREC)*, 2020, pp. 6063–6068.
[55] M. Kaczmarek and M. Filhol, "Alignment data base for a sign language concordancer," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 6069–6072.
[56] M. Mukushev, A. Sabyrov, A. Imashev, K. Koishibay, V. Kimmelman, and A. Sandygulova, "Evaluation of manual and non-manual components for sign language recognition," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1–6.
[57] F. Zhang, S. Han, H. Gao, and T. Wang, "A Gaussian mixture based hidden Markov model for motion recognition with 3D vision device," *Comput. Electr. Eng.*, vol. 83, May 2020, Art. no. 106603.
[58] S. Paris, "Fast GMM and Fisher vectors," MathWorks, Natick, MA, USA, Tech. Rep. 38372, 2021.
[59] D. Avola, L. Cinque, M. De Marsico, A. Fagioli, and G. L. Foresti, "LieToMe: Preliminary study on hand gestures for deception detection via Fisher-LSTM," *Pattern Recognit. Lett.*, vol. 138, pp. 455–461, Oct. 2020.
[60] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3412–3419.
[61] C. K. M. Lee, K. K. H. Ng, C.-H. Chen, H. C. W. Lau, S. Y. Chung, and T. Tsoi, "American sign language recognition and training method with recurrent neural network," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114403.
[62] R. Rastgoo, K. Kiani, and S. Escalera, "Real-time isolated hand sign language recognition using deep networks and SVD," *J. Ambient Intell. Hum. Comput.*, vol. 13, pp. 591–611, 2022.
[63] S. Y. Boulahia, E. Anquetil, F. Multon, and R. Kulpa, "Dynamic hand gesture recognition based on 3D pattern assembled trajectories," in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2017, pp. 1–6.
[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 *Proc. 10th Eurograph. Workshop 3D Object Retr. (3DOR)*, 2017, pp. 1–6.
[65] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006.
[66] J. Nocedal and S. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 2006.
[67] J. Ngiam, Z. Chen, S. Bhaskar, P. Koh, and A. Ng, "Sparse filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1125–1133.
[68] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1–9.
[69] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 465–470.
[70] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 716–723.
[71] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1340–1352, Jul. 2015.
[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 *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 273–286.
[73] J. Liu, Y. Liu, Y. Wang, V. Prinet, S. Xiang, and C. Pan, "Decoupled representation learning for skeleton-based gesture recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5751–5760.
[74] K. Lupinetti, A. Ranieri, F. Giannini, and M. Monti, "3D dynamic hand gestures recognition using the leap motion sensor and convolutional neural networks," in *Proc. Int. Conf. Augmented Reality, Virtual Reality Comput. Graph.* Cham, Switzerland: Springer, 2020, pp. 420–439.
[75] B. Hisham and A. Hamouda, "Arabic sign language recognition using adaboosting based on a leap motion controller," *Int. J. Inf. Technol.*, vol. 13, no. 3, pp. 1221–1234, Jun. 2021.

SUNUSI BALA ABDULLAHI (Member, IEEE) received the B.Sc. and M.Sc. degrees in electronics from Bayero University Kano (BUK), Nigeria. He is currently pursuing the Ph.D. degree with the King Mongkut's University of Technology Thonburi, Thailand. His current research interests include computer vision, artificial intelligence, nonlinear optimization and their applications in human motion analysis, data analysis, wireless systems, and social signal processing.

KOSIN CHAMNONGTHAI (Senior Member, IEEE) received the B.Eng. degree in applied electronics from The University of Electro-Communications, in 1985, the M.Eng. degree in electrical engineering from the Nippon Institute of Technology, in 1987, and the Ph.D. degree in electrical engineering from Keio University, in 1991. He is currently a Professor with the Department of Electronic and Telecommunication Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi. His research interests include computer vision, image processing, robot vision, signal processing, and pattern recognition. He is a member of IEICE, TESA, ECTI, AIAT, APSIPA, TRS, and EEAAT. He is the Vice President-Conference of APSIPA Association (2020–2021). He has served as the Chairperson for the IEEE COMSOC Thailand, from 2004 to 2007, and the President for ECTI Association, from 2018 to 2019. He has served as an Editor for *ECTI E-Magazine*, from 2011 to 2015, and an Associate Editor for *ECTI-EEC Transactions*, from 2003 to 2010, and *ECTI-CIT Transactions*, from 2011 to 2016.