IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 November 2022, accepted 12 December 2022, date of publication 2 January 2023, date of current version 9 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3233671

RESEARCH ARTICLE

Multi-Semantic Discriminative Feature Learning for Sign Gesture Recognition Using Hybrid Deep Neural Architecture

E. RAJALAKSHMI[1], R. ELAKKIYA[1,2], V. SUBRAMANIYASWAMY[1], L. PRIKHODKO ALEXEY[3], GRIF MIKHAIL[3], MAXIM BAKAEV[3], KETAN KOTECHA[4], LUBNA ABDELKAREIM GABRALLA[5], AND AJITH ABRAHAM[6], (Senior Member, IEEE)

[1]School of Computing, SASTRA Deemed University, Thanjavur, Tamil Nadu 613401, India
[2]Department of Computer Science, Birla Institute of Technology and Science, Pilani, Dubai Campus, Dubai International Academic City, Dubai, United Arab Emirates
[3]Department of Automated Control Systems, Novosibirsk State Technical University, 630073 Novosibirsk, Russian
[4]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 412115, India
[5]Department of Computer Science and Information Technology, College of Applied, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[6]Faculty of Computing and Data Sciences, FLAME University, Pune 412115, India

Corresponding authors: R. Elakkiya (elakkiyaceg@gmail.com) and V. Subramaniyaswamy (vsubramaniyaswamy@gmail.com)

The financial support for the publication is done by Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R178), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT The speech and hearing-impaired community use sign language as the primary means of communication. It is quite challenging for the general population to interpret or learn sign language completely. A sign language recognition system must be designed and developed to address this communication barrier. Most current sign language recognition systems rely on wearable sensors, keeping the recognition system unaffordable for most individuals. Moreover, the existing vision-based sign recognition frameworks do not consider all of the spatial and temporal information required for accurate recognition. A novel vison-based hybrid deep neural net methodology is proposed in this study for recognizing Indian and Russian sign gestures. The proposed framework is aimed to establish a single framework for tracking and extracting multi-semantic properties, such as non-manual components and manual co-articulations. Furthermore, spatial feature extraction from the sign gestures is deployed using a 3D deep neural net with atrous convolutions. The temporal and sequential feature extraction is carried out by employing attention-based Bi-LSTM. In addition, the distinguished abstract feature extraction is done using the modified autoencoders. The discriminative feature extraction for differentiating the sign gestures from unwanted transition gestures is done by leveraging the hybrid attention module. The experimentation of the proposed model has been carried out on the novel multi-signer Indo-Russian sign language dataset. The proposed sign language recognition framework with hybrid neural net yields better results than other state-of-the-art frameworks.

INDEX TERMS Indian sign language recognition, isolated sign language recognition, deep neural network, multi-semantic sign features, attention mechanism, gesture recognition, sign language.

I. INTRODUCTION

Sign Language (SL) is the basic means of interaction for the speech-impaired and hard-of-hearing populace. Like any other language, Sign Language seems to have its underlying structure and grammatical rules that allow users to communicate and express themselves adequately. Furthermore, the SL is usually expressed through manual components such as Hand motion, Hand position and non-manual articulations such as eye gaze, facial expression, lip movement, etc. The manual and non-manual components together form the multi-semantic feature components. Mastering an SL requires substantial effort for the hearing community, which demands developing a Sign Language Recognition

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

system (SLR). Recently developed Sign Language datasets include American, Arabic [1], German, Chinese [2], Turkish [3], Bhutanese [4], Russian, and Indian Sign languages (ISL) [5]. SLR has been intensively investigated to help hearing people comprehend sign language and make the everyday lives of the speech-impaired and hard-of-hearing community more convenient. The SLR frameworks aim to detect and recognize sign language performed by the sign interpreter from a visual medium. Various concerns in developing an SLR include signer-dependent variations, local and global elements, feature extrication from heterogeneous backgrounds, large vocabulary and scalability, multi-modality, occlusion, and movement epenthesis. Although countless research on SLR has been undertaken, most issues remain unresolved. Most of the developed SLRs leverage wearable sensors, colour-coded gloves, or multiple depth sensor cameras to capture the data, making the signer very uncomfortable conveying the sign gestures in real-life scenarios.

Moreover, deploying such sensors makes the recognition system very costly and unaffordable for common people. Although few SLR frameworks have been established, the abovementioned concerns couldn't be resolved completely without leveraging external sensors. Most of the existing SLRs impose some signer clothing restrictions to avoid the complex background issues completely, while some didn't address all the multi-semantic feature learning. Moreover, occlusion and movement epenthesis complications still plague the majority of SLR frameworks.

A novel vision-based Multilingual Sign Language Recognition framework is proposed to track and extract multi-semantic, manual co-articulations, including one-handed and double-handed signs and non-manual elements, including facial expressions, body, and lip movement. The fundamental motivation for designing the proposed work is to promote a more natural style of signing with no clothing restrictions, eradicate the use of wearable devices or sensors or gloves, handle manual features such as one-handed and two-handed signs and non-manual features, including facial expression, body, and lip movement, and develop a unified framework for multilingual Sign Language Recognition. The proposed architecture intends to learn the multi-semantic features by implementing two modules in Phase-I. The Manual Articulation Tracking Module's first module helps detect the manual components from the full-frame sign images. The Non-manual Element Tracking module's second module helps extract the signer facial features from the sign frames. The extracted features are then learned using a hybrid Deep Neural Net (hDNN) framework, including a 3-D neural net with atrous convolution and Attention-based Bidirectional LSTM (Bi-LSTM), for extraction of spatial, sequential, and temporal components from the sign gestures. In Phase II, the discriminative features are selected in the Discriminative Feature Detection module. The main contributions of the proposed hDNN approach are as mentioned below:

1. Create a novel, natural, multi-signer Indo-Russian Sign Language Database comprising isolated sign gestures.
2. Extract and select multi-semantic manual one-handed and two-handed signs and Non-manual gestures such as body movement and expressions
3. Develop hybrid deep 3-D Neural Net and Bi-LSTM to extract spatial, sequential, and temporal feature components. Using hybrid Autoencoders, extract more abstract characteristics. Extract and select Discriminative features using hybrid attention.
4. Track and Recognize Isolated sign gestures using the novel Indo-Russian Sign Language database and evaluate the outcomes and performances using other state-of-the-art approaches.

This article is split up into different segments. The review of the literature and related works are covered in Section II. Section III explains the Proposed Approach. The outcomes and discussion are detailed in Section IV, and the concluding remarks and future works are addressed in Section V.

II. RELATED WORK

Many investigations have been performed in the realm of Sign Language Recognition. Various researchers have already published new variations for establishing the sign structure in the perception of subunits analogous to phonology in a spoken language [6], [7]. The SL's components can be defined using the Stokoe paradigm depending on the motion, alignment, and shape [8]. The fingers' layout and the palms' orientation indicate the hand shape, whereas the hand positioning about the frame is defined. The Movement-Hold model is typically composed of sequential [47] movement organization and static posture that is continual signing [9]. Although the relationship between the motion and stances of the hands has indeed been approximated in two-handed signs, the approaches mentioned above are sufficient for one-handed signs [10]. Vision-based (multi-semantic) and sensor-based (multi-modality) are different SLRs. Physical sensors, such as infrared depth and maps [13], [14], were employed in the early efforts to attain multi-modality [11], [12] for acquiring 3-D spatial intelligence. Researchers have explored the integration of RGB and optical flow in a handful of works, resulting in better outcomes on the PHOENIX-14 collection. The development of an SLR poses a variety of obstacles. Even though the same individual repeatedly gestures the same sign, minor hand pace and location variations are noticed.

The major challenges in the SLR include hand segmentation and tracking from various contexts and environments, occlusion, illumination variation, hand orientation, etc. [15]. The isolated SLR research has lessened the barrier of tracking and segmentation by directly detecting location features with devices on hands, such as coloured mittens or markings. An alternative paradigm was presented, including Sign instructor steps for recognizing, evaluating, and categorizing hands and faces [16]. Coloured gloves have been employed in this project to make hand identification and segmentation easier. A similar methodology was introduced using colour-coded mittens, making human hand tracking easier. Techniques used a multi-colour detection approach and an HMM

for categorizing significant attributes from recordings. The hand was considered the single identified item in a multi-function extraction [17]. Some other authors presented a directed histogram and classifications employing Euclidean Distance and KNN to discern static, isolated signs with a 90% accuracy [18]. An SLR with an Optimized Neural Net was devised [19]. A similar model was developed using the open pose to derive the posture cues of the Signing gestures [20]. In [21], different SVM and skin segment-based methods were formulated.

An SLR was given a special twist when a sophisticated background setting was combined with dynamic sign gestures [22]. One of several projected works [23] also included a dynamic SLR system for ISL. A grid-based system was presented employing ISL [24] for real-time hand position and sign recognition. An alternative Isolated Sign Language Recognition (ISLR) with Bayesian KNN was deployed for eliminating sign redundancies [15]. Some researchers focused on differentiating between manual and non-manual elements [25]. A word-level pose-based SLR also was devised employing BERT and GCN [26]. Leveraging Neural Machine Translation, an SL translation system was constructed for end-to-end and pre-trained contexts [27]. In a previous study, the researchers introduced a time-based accumulative element detection for building ISLR. To retrieve the manual aspects of the signing motions, a Convolution Neural Network was used to synthesize modelling-based hand energy ISLR [28]. DeepArSLR, a robust ISLR method, was developed to handle multi-signer constraints in Arabic SL leveraging Deep Learning [29]. An ISLR with a word dataset was developed in which the scheme learned domain invariant descriptors by transferring knowledge of headline gestural captions [30]. The attention mechanism model addressed the complex background constraint for gesture detection [34]. An ISLR has been accomplished by utilizing a neural methodology deployed using CNN [33]. The SL Graph Convolution Network [32] has presented a dynamic sign recognition. Spatio-temporal video-based ISLR with the Deep Cascade model [31] was used in other approaches. A deep neural architecture with CNN+Stacked LSTM [35] was built to accommodate static and dynamic responses. Extensive research has also been conducted on several regional SL. For Sinhala word numeric sign detection, a Sinhala SLR system was developed [36]. A database was collected for Arabic SLR also [37]. Several implementations of SLR have been built for the Indian SL system [38], [39]. Eventually, an SLR framework was developed applying syntactically directed Korean Sign Language recognition [40]. Various image processing techniques [46] are also available for Sign Language Recognition.

III. PROPOSED METHODOLOGY

While examining a sign representation video for dynamic motions, $v_s = \{v_s\}_{f=1}^{F}$ comprising F frames, the main goal is to determine the relevant sign class, w_l. The hDNN is signer independent and has been shown to handle concerns such as varied complexion, illumination changes, complicated backgrounds, facial expressions, and hand sizes, among others. hDNN is developed to identify the Isolated Word Sign Gestures. The framework mainly aims at discriminative feature learning from isolated sign gestures.

A. DATA PREPARATION

For sign recognition, the videos were transformed into frames. Depending on the duration of the sign activity, each video generates a varying number of frames. So here, the Frame Sampling (FS) technique has been used to cope with these inconsistencies and bring uniformity to the series of frame images produced by all sign videos. The FS approach aids in extracting a set of predetermined frame images from sign gesture videos. In FS, the average of the number of frame images of every video, f_{avg}, is estimated and kept as the threshold count and the f_{avg} is then set up as the predetermined frame counts for each video. The f_{avg}, can be computed as given in Eq. 1 where C_f indicates the total count of frame images from all the sign videos and C_v indicates the total count of videos.

$$f_{avg} = \frac{C_f}{C_v} \tag{1}$$

The threshold value chosen is f_{avg}, therefore the first f_{avg} frames from each video are retrieved for further processing. Whenever the frame count in a sign video falls below the f_{avg}, the very last image is duplicated for having the number of frames for each video consistent as f_{avg}. In this manner, the number of frames produced by all the videos will get set to a constant value i.e. f_{avg}. Hence all the video clips will have a consistent number of frames.

B. PROPOSED SLR FRAMEWORK WITH HYBRID DEEP NEURAL NET

The proposed hybrid Deep Neural Net framework intends to recognize multilingual, multi-semantic signer independent Sign Language recognition (hDNN-SLR). The hDNN-SLR framework deals with the semantic manual co-articulation and non-manual element detection and spatial, sequential and temporal component extraction. Figure 1 illustrates the proposed hDNN-SLR framework. Since there is a lack of ISL and Russian SL (RSL.) Database, the first step toward recognition is the creation of a novel multi-signer SL dataset. The SL dataset was created for isolated sign gestures. The next step deals with the dataset preprocessing module wherein the SL videos were prepared for recognition. In this module, the Sign videos were first converted into frame images. Then, all the frame images were resized, and frame sampling was done to bring uniformity and consistency. Data Augmentation was done to increase the SL dataset to have better training. The preprocessed frame images were then provided as input for the Multi-semantic component Extraction module for non-manual and manual element tracking from the video frame images.

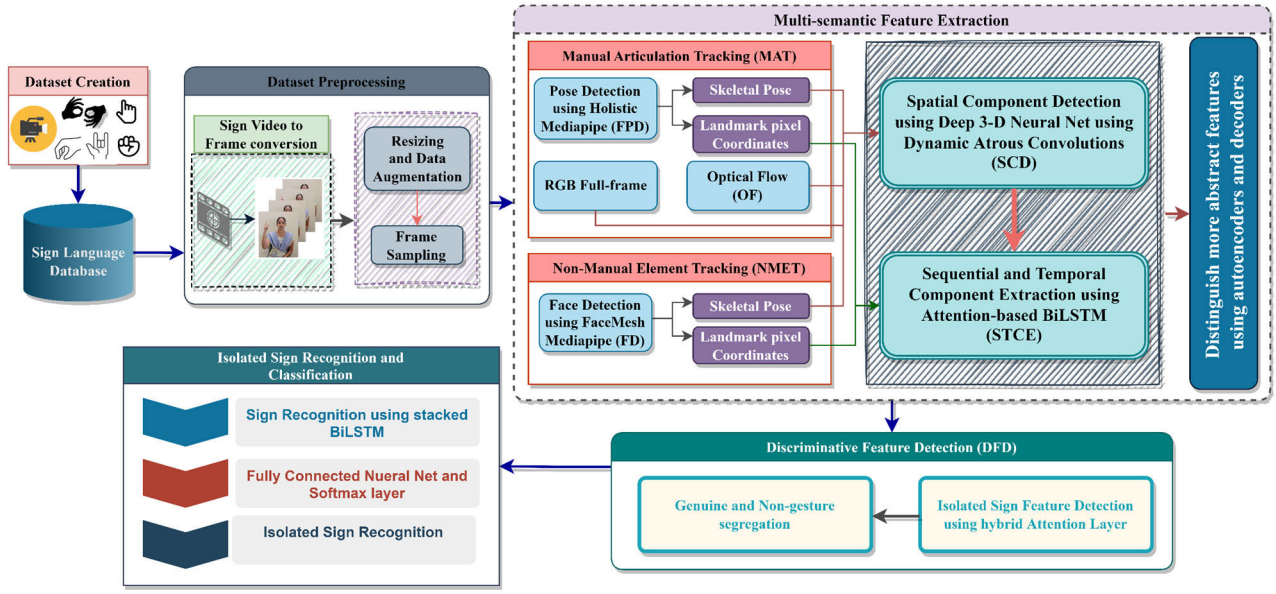

FIGURE 1. Proposed hDNN-SLR framework.

The Multi-semantic Feature Extraction unit is divided into submodules, namely Manual Articulation Tracking (MAT) and Non-Manual Element Tracking (NMET) for feature detection and Spatial Component Detection module (SCD), Sequential-Temporal Component Extraction module (STCE) for feature learning. In the MAT submodule, the tracking of semantic manual articulations is carried out. In this module, Holistic MediaPipe tracks hand, palm, and full pose skeletal from the sign frame images. The Holistic Mediapipe generates skeletal pose and Landmark pixel coordinates from the sign gestures. The skeletal posture is then given input to the SCD module to extract spatial features. The landmark coordinates generated by the Holistic MediaPipe are sent to STCE Component to extract the temporal and sequential components.

Furthermore, in the MAT module, Full RGB frames of the sign videos are also used for feature extraction to overcome the issues posed by Holistic MediaPipe. Optical Flow (OF) is generated from the RGB sign frame images for tracking the velocity of the motion of gestures, thereby analyzing the sign articulations in depth. The RGB and Optical flow frame images are input to the SCD module for spatial feature extraction. The next module is the NMET module, wherein the Non-Manual elements are detected and tracked from the sign images. FaceMesh MediaPipe is leveraged to extract the facial elements from the sign gestures in this module. The skeletal face pose generated by the Face Detection module (FD) is given as input to the SCD module for spatial feature extraction. In contrast, the facial Landmark pixel coordinates generated are inputted into the STCE module.

The SCD module gets the frame image inputs from the MAT and NMET modules. The SCD module helps to extract the spatial features from the sign gesture video frames by leveraging a Deep 3-D Neural Net using Dynamic Atrous Convolutions (DAC). The Feature Maps generated by the SCD are then given as input to the SCTE module to extract the temporal and sequential information from the sign gestures using Attention-based Bi-LSTM (A-BiLSTM). The SCTE module receives inputs from MAT, NMET, and SCD modules. The output from the SCTE module is further passed to the autoencoders and decoders for extracting the distinguished abstract features. After the feature extraction from the Multi-semantic Feature Learning module, the extracted characteristics are given as input to the Discriminative Feature Learning module. In this module, discriminative features are extracted using a hybrid Attention Layer that helps to recognize the genuine and non-genuine features. The feature extracted are then given to the Sign Recognition and Classification module for recognition of Isolated Sign Language.

1) MANUAL ARTICULATION TRACKING

The manual semantic co-articulations comprise hand shape, palm and finger orientation, movement, etc. The manual elements and full pose are extracted in the Manual Articulation Tracking (MAT) module. The MAT module comprises Pose Detection using Holistic MediaPipe, RGB Full-frame feature detection, and Optical flow component tracking. The full-body pose, hand orientation, and movements are tracked and estimated using the MediaPipe Holistic pipeline in the Full Pose Detection submodule (FPD). Google created a cross-platform library called Mediapipe that offers incredible pre-built machine learning (ML) solutions for computer vision problems. For typical tasks like hand tracking and pose tracking, MediaPipe offers foundational machine learning models, addressing the same barrier in development that affects various machine learning applications. Separate

models for posture, face, and hand components are individually integrated into the MediaPipe Holistic pipeline and optimized for their respective fields. However, the input to one component is not appropriate for the others due to their disparate areas of expertise. For instance, a smaller, fixed-resolution video frame (256×256) is used as input for the posture estimation model. But the picture quality would be too low for precise articulation if the hand and facial areas were cropped out of that image and given to the appropriate models. As a result, MediaPipe Holisticis created a multi-stage pipeline that processes the various regions using an image resolution suited for each one. FPD combines independent models for the posture, face, and hand components, each optimized for its realm. While reacting to quick gestures, FPD utilizes posture prediction (on each frame) as an auxiliary ROI to minimize the pipeline's latency beforehand. Hence, the framework preserves semantic consistency by limiting a mix-up between the right and left hands or parts of the body. The pose generated, tracing the sign gesture in the frame, using FPD, is then given as input to the Spatial Component Detection (SCD) for spatial feature learning from the full pose generated.

Figure 2 shows the traced skeletal pose generated in the FPD module. The FPD also detects a total of 543 landmark coordinates of a face (468), hand (21), and pose (33) landmark coordinates along with the skeletal pose. These landmark coordinate values are then fed to the Sequential and Temporal Component Extraction (STCE) Module for further feature learning. The Posture data perform poorly when the hands face parallel to the floor/ceiling or perpendicular to the camera. Hence, RGB Full-frame feature extraction and Optical Flow (OF) component tracking is deployed along with the skeletal pose estimation to overcome such issues and for efficient tracking of the hand gesture movements. The RGB Full-frame images are directly passed to the SCD module for spatial component feature extraction. Optical flow is a term used to track the motion in the frame. Optical flow detects the velocity of points inside frames and predicts where points might appear in the future frame image series. Optical flow information is derived from successive images to comprehend the dynamic movement of signers. M optical flow images are inputted into the stream for optical flow information using the TVL1 method [49], which is calculated from $M + 1$ successive frame images. Hence the optical flow from the Full frame is extracted and passed to the SCD module.

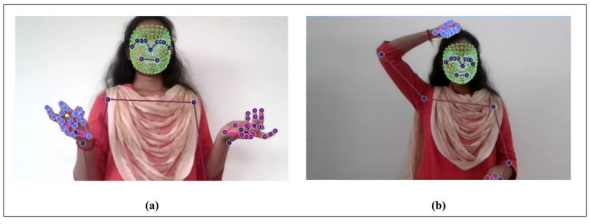

FIGURE 2. Sign skeletal pose generated by FPD module.

2) NON-MANUAL ELEMENT TRACKING

The Non-manual components comprise Facial expression, lip movement, eye gaze, etc. The proposed framework detects and extracts the semantic Non-manual feature components with the help of the Non-Manual Element Tracking (NMET) Module. The NMET comprises two modules: Face Detection using the FaceMesh module, FD, and Mouthing Cue Detection module, MCD. The FD module leverages the MediaPipe FaceMesh pipeline for extracting the facial components. The Facial Transform module of FaceMesh fills the gap between accurate real-time augmented reality applications and face landmark estimates. It creates a metric 3D space and then estimates a face transformation inside it using the locations of the facial landmark screens. The face transform data comprises typical 3D primitives such as a triangular face mesh and a face position transformation matrix. The FD module generates a real-time facial geometry framework that detects 468 3D facial landmark coordinates with the help of the FaceMesh pipeline. It uses computer vision to deduce 3D surface geometric features from regular camera inputs, eliminating a specialized depth sensor requirement. FD module also leverages an approach based on attention with FaceMesh pipeline to semantically relevant facial regions, anticipating landmarks quite efficiently around irises, eyes, and lips at the cost of more computation and the face landmark framework. The landmark coordinates are then inputted to the STCE module, and the facial skeletal pose traced and generated is inputted to the SCD module for spatial feature extraction. Mouth shape, eye gazing, and facial expressions are common fine-grained cues in SL articulation. These cues fluctuate with time and cover small spatial regions in the Spatio-temporal pipes. As a result, they quickly vanish in subsequent convolution and pooling. To sustain those fine-grained elements, we employ an additional adaptive sampling to provide a Facial Cue Enhancement (FCE) module highlighting them with increased resolution.

3) SPATIAL COMPONENT DETECTION

The spatial features from the sign video frames are extracted using a Spatial Component Detection (SCD) module that leverages the Deep 3-D Neural Nets with Dynamic Atrous Convolution (DAC). 3-D conv Neural Net employ a 3-D filter on the input, traversing across three directions, to obtain the low-level feature embedding. Its output is having a 3-D volume space. Dilated or Atrous convolution define the space between units in filters. In this convolution type, the kernels' receptive view extends due to the space; for example, the visual field acquired by the 3×3 kernels with a dilation rate of 2 is identical to the field of vision acquired by the 5×5 kernel. The complexity maintains the same in this circumstance, but distinct features are produced. The DAC helps accelerate the feature extraction process for many sign gesture frames. The feature map generated by combining the temporal feature, t, can be represented as $M \in \mathbb{R}^{t \times h \times w}$. Given the coordinate points $p = (p_t, p_x, p_y)$ The sampling can be

denoted by $s \in \mathbb{R}^{k\times 3}$. So we can have a $3\times3\times3$ convolutional operation generating $k = 27$ entries. So we have dilated map, $d \in \mathbb{R}^{t\times h\times w\times 3}$. For the prediction of d, an input feature map x is inputted to a basic $3\times3\times3$ convolution d_f. After generating the feature map, we apply an addition operation by 1 with an elu activation function, as illustrated in Eq. 2.

$$d = 1 + elu(d_f(x)) \tag{2}$$

Essentially constrains the elements in d to the range $[0, \infty)$. It was discovered that just engaging $d_f(x)$ with a ReLU activation eventually ends up in zero gradients. The Eq. 2 formulation, on the other hand, yields far better gradients throughout learning and superior inference outcomes. Parallel to this, to generate a modulation map $m \in \mathbb{R}^{t\times h\times w\times k}$, a $3 \times 3 \times 3$ convolution m_f is adopted into x along with an activation function, namely sigmoid. With coordinate p_0, using DAC, the output y is computed as illustrated in Eq. 3 where $(p_i \cdot d(p_0)$ indicates the multiplication of the sampling coordinate points of the location p_i and tuple with a dilation rate of 3 in d at p_0.

$$y(p_0) = \sum_{p_i\in s} m(p_0, p_i) \cdot w(p_i) \cdot x(p_0 + (p_i \cdot d(p_0))) \tag{3}$$

In each convolutional layer, the Dilation and Modulation outcomes were concatenated. The dilation rates vary according to the position of the signer inside the frame images. The 3-D Deep Neural Architecture is detailed in Table 1. Cross-Entropy having a batch size of 128 with a learning rate of 1, was employed for training. Figure 3 illustrates the framework deployed for spatial, temporal and sequential feature learning with SCD and STCE modules.

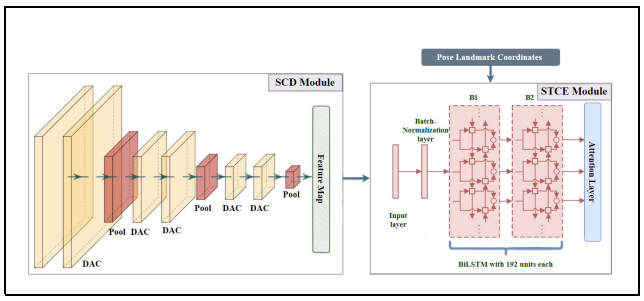

FIGURE 3. Spatial, temporal and sequential feature extraction using SCD and STCE module.

4) TEMPORAL AND SEQUENTIAL COMPONENT LEARNING

The temporal and sequential feature learning from the sign frame images is done in the Sequential and Temporal Component Extraction Module (STCE) by employing the Attention-based Bi-LSTM (A-BiLSTM). The STCE module is divided into 2 phases: Phase I for deploying A-BiLSTM for analyzing MediaPipe-generated skeletal landmark coordinates and Phase II for deploying A-BiLSTM for analyzing the feature maps generated by the 3D Neural Net. In both Phases, the underlying architecture is the same, while the input and the outcome may differ accordingly. The input layer is the 1^{st} layer of A-BiLSTM. In Phase-I of A-BiLSTM, the input layer is the matrix form representing the x, y, z coordinates of the skeletal pose generated by MediPipe. In contrast, in Phase II of A-BiLSTM, the input layer is the matrix form representing the t, h, w as the temporal feature, height, and width attributes. The Batch Normalization (BN) layer is the second layer. Using a higher learning rate, the BN approach [36] can cut down the number of learning steps involved in network convergence without paying much attention to dropouts and baseline parameters. As a result, a BN is utilized to streamline and automate the Network's training. The 2^{nd} and 3^{rd}, namely, $B1$ and $B2$, deploy Bi-LSTM, each having 192 units. The input of $B2$ layer is the sequential outcome of the $B1$ layer. Two Bi-LSTM layers are added after the BN layer since experiments have already established [37] that over two recurrent layers are more effective in forecasting temporal activities. The activation function utilized in this Network is the Tanh function. Since the Tanh function's output is -1 to 1, which corresponds to the feature distributions of most sequences focused on 0, it seems to have a higher gradient than the Sigmoid function at the insight of 0, and the model can indeed be faster to convergence. L2 emits the hidden unit elements of all epochs (A1). The reverse and forward LSTM make up Bi-LSTM, whereas the Input gate, Forget gate and memory cell make up an LSTM memory block. The computations involved in the Bi-LSTM are described in Eq. 4 to Eq. 2. In the given equations, F_t denotes the forget gate output, I_t denotes the input gate outcome, O_t denotes the outcome of the output gate, $\tilde{M}_t$ denotes the cell's candidate outcome while M_t denotes the state of the cell. The outcome of the memory cell is denoted by B_t, the outcome of the forward and reverse LSTM is denoted by f_h and r_h. The weight and bias entities are represented as w and b.

TABLE 1. 3D deep neural net architecture with DAC.

Layers	Filter	Stride	Dilation Rate	Feature Map
DAC	$3\times3\times3$	$1\times1\times1$	$3\times3\times3$	4
DAC	$3\times3\times3$	$1\times1\times1$	$3\times3\times3$	4
Max-Pooling	$4\times4\times4$	$2\times2\times2$		4
DAC	$3\times3\times3$	$1\times1\times1$	$3\times3\times3$	8
DAC	$3\times3\times3$	$1\times1\times1$	$3\times3\times3$	8
Max-Pooling	$4\times4\times4$	$2\times2\times2$		8
DAC	$3\times3\times3$	$1\times1\times1$	$3\times3\times3$	16
DAC	$3\times3\times3$	$1\times1\times1$	$3\times3\times3$	16
Max-Pooling	$4\times4\times4$	$2\times2\times2$		16

$$F_t = \sigma(W_F \left[H_{(t-1)}, X_t\right] + b_F) \tag{4}$$
$$I_t = \sigma\left(W_I \left[H_{(t-1)}, X_t\right] + b_I\right) \tag{5}$$
$$\tilde{M}_t = \tanh\left(w_m * \left[H_{(t-1)}, X_t\right] + b_C\right) \tag{6}$$
$$M_t = F_t M_{(t-1)} + I_t * \tilde{M}_t) \tag{7}$$
$$O_t = \sigma\left(w_o * \left[H_{(t-1)}, X_t\right] + b_O\right) \tag{8}$$
$$H_t = O_t * \tanh\left(M_t\right) \tag{9}$$

$$f_h = f\left(w_{f1}X_t + w_{f2}H_{(t-1)}\right) \quad (10)$$
$$r_h = f\left(w_{r1}X_t + w_{r2}H_{(t+1)}\right) \quad (11)$$
$$B_t = g\left(w_O * f_h + w_O * r_h\right) \quad (12)$$

The Attention layer is employed based on the significance of the sign gesture's sequential and temporal essence. Eq. 13 to Eq. 15 illustrates the integration of the Attention layer to Bi-LSTM to deploy A-BiLSTM architecture where i_t denotes the unit of the hidden net, B_t is considered as the outcome of BiLSTM, the weight coefficient vector is represented by a_t, w_u represents the weighted co-efficient matrices from $B2$ to the Attention layer, j_t represents the outcome of the Attention layer, and b represents the bias entity. In A-BiLSTM, the vector i_w is utilized to extract the temporal aspects.

$$i_t = \tanh(w_u B_t + b) \quad (13)$$
$$a_t = softmax\left(i_t^T i_w\right) \quad (14)$$
$$j_t = \sum a_t B_t \quad (15)$$

The dot product between i_t and i_w, helps to measure the significance of characteristics. The value of a_t is estimated with the help of the softmax activation function. The temporal attention mechanism imparts varying weights to various aspects of sign gestures at distinct intervals, so significant aspects get greater attention, thereby improving gesture recognition accuracy.

5) DISTINGUISHED ABSTRACT AND DISCRIMINATIVE FEATURE DETECTION

Our framework used Variational Auto-Encoders (VA-E) to learn distinguished abstract components from the sign gestures, thereby removing the noise elements. Fingers, palm radius, palm orientation, hand orientation, the separation between fingertips and palm, and palm radius are among the high-level parameters extracted utilizing the VA-E. The VA-E is a regularized auto-encoder that guarantees that the latent space has acceptable parameters for generative processes and eliminates overfitting. Rather than a single point, VA-E captures the knowledge as a dispersion all through the feature space, yielding some latent space regularization. Provided a base underlying probabilistic framework to characterize the feature elements, the loss function of VA-E, consisting of a reconstruction component and a regularization component, might well be meticulously constructed by employing the statistical model of variational inference. The loss function has been greatly reduced while training a VA-E. The reconstruction component makes the encoding and decoding design efficient. The regularization component on the residual layer attempts to regulate the residual space organization by keeping the encoder distribution closer to a conventional normal distribution. The regularisation component is the Kulback-Leibler divergence between the conventional Gaussian and the returned distributions.

The Dropout generalizes the Neural Net structure by performing multiplication on various activations with 0 at random. In contrast, ReLU activation has enabled Neural Nets to achieve convergence faster and better than the other standard activation functions. The output of a neuron can be determined using either of these methods. Regardless, the two operate independently of one another. Furthermore, Zone out regularizer multiplies the input with 1 in a stochastic method. Hence to bring these three together, along with the VA-E, GELU [41] activation function is employed. It helps to have the result of the activation function deterministically by incorporating three functions and stochastically performing multiplication on input by 1 or 0. With the modified VA-E's help, we can extract high-level abstract features from the sign gestures.

After abstract feature learning, the extracted components are given to the Discriminative Feature Detection Module (DFD). The DFD module comprises of hybrid Attention mechanism (h-Attention). The h-Attention layer integrates the Segmentation and Spatial attention along with Light-BGM. The attention mechanism's task is to provide context and direct the attention of the decoder to specific encoder range outcomes. The ultimate goal of attention mechanisms is to produce a weighted representation of the source sequence to assist in decoding. The terminology context vector describes this summarization. Let $U = [u_1, u_2..u_{z+1}]$ be the hidden components, considering o_n as each state component's output, a_n and c_n as the attention weight and context vector, respectively, we can formulate the c_n as given in Eq. 17.

$$a_n = softmax(o_n WU) \quad (16)$$
$$c_n = [a_n U_n] \quad (17)$$

But, using this Attention method, it would not be easy to distinguish between the weights of different semantic cues. Hence h-Attention has been introduced that integrates Spatial Attention (hSA) and Segmentation Attention (hSE). The hSA generates a spatial attention map by exploiting the inter-spatial interconnectedness of features. In contrast to channel attention, where the focus is on channel placement, hSA is complementary to channel attention and focuses on the placement of the informational component. hSA convolves average-pool and max-pool layers with the channel vectors to provide a meaningful feature descriptor for quantifying spatial attention. The input has been taken from the SCD output feature maps. To generate the hSA feature maps, $S_m(f) \in \mathbb{R}^{h\times w}$. S_m determines the components that are needed to be emphasized. The channel information from the feature maps is incorporated with *two* pooling operations, thereby creating $2D$ feature maps. The 2D feature maps $M_s^{avg} \in \mathbb{R}^{1\times h\times w}$ and $M_s^{max} \in \mathbb{R}^{1\times h\times w}$ represent the corresponding average and max pooling. By convolving and concatenating the 2D feature maps generated, the hSA map is produced. The hSA can be formulated as illustrated in Eq. 16. The σ notation denotes the sigmoid function and $F^{7\times7}$ denotes the convolution operation with a filter size of 7×7. The final feature map was flattened and concatenated with the hSE.

$$S_m(f) = \sigma(F^{7\times7}([Avgpool\,(f)\,; Maxpool(f)])) \quad (18)$$

$$S_m(f) = \sigma(F^{7\times7}([M_s^{avg}, M_s^{max}])) \quad (19)$$

The attention weights are assigned to each feature extracted independently by the hSE method. The c_n can be computed using hSE with the help of Eq. 22. Here a_n is segmented uniformly across $z+1$ channel.

$$o_n = [o_{n,1}, o_{n,2}, \ldots, o_{n,z+1}] \quad (20)$$

$$a_n = softmax(o_{n,z}W_zU_z) \quad (21)$$

$$c_n = [a_{n,1}, u_1, a_{n,2}, u_2, \ldots, a_{n,z+1}, h_{z+1}] \quad (22)$$

The hAttention mechanisms help to segregate the useful genuine gestures from the non-useful gestures, thereby reducing the computation for sign recognition. The output of the h-Attention is then passed to the sign recognition task. The sign recognition is done using the stacked Bi-LSTM by stacking three layers of Bi-LSTM, thereby forming a Bi-LSTMNet. The Bi-LSTMNet consists of varying hidden layers for efficient recognition. The Bi-LSTM unit is the recurrent unit for its capacity to handle long-term dependencies. Bidirectional inputs are combined into forwarding and backward hidden states by the BLSTM. To produce spoken language translations, the hidden state of each time step is transmitted via a fully-connected layer and a softmax layer.

IV. RESULTS AND DISCUSSIONS

A. DATASET

The lack of availability of the ISL and RSL datasets has mandated the creation of a novel Sign Language database. The proposed research involves collecting and creating a novel, multi-signer Indian-Russian Isolated SL dataset. For the Indian Isolated Word Sign dataset (IIWS), sign videos were captured using a DSLR camera with 30-45fps. The sign videos were created with the participation of seven volunteer signers. Three female volunteers, ages 26, 23, and 25, and four male volunteers, ages 24, 27, 30, and 33, participated as volunteers for the IIWS sign gesture database collection. While collecting the dataset, no clothing restrictions were considered, and none of the wearable sensors or devices was utilized. The videos were taken under varying illumination conditions. IIWS comprises random 500-word sign gestures that are used in our daily lives. At least five repetitions from each signer have been considered for each word. After applying Data Augmentation, the IIWS dataset consists of 3000-word sign videos (mp4), each having 1080p resolution. All the videos were transformed into frame images in JPEG format. Individuals with diverse complexion and hand sizes were selected to acquire the data. The video clips were shot in a closed environment with standard lighting.

Nearly 1100 unique, commonly used Russian word sign videos have been created and collected regarding the Russian dataset. A dataset comprising 1100 unique signs and 37775 standard Russian sign representations was generated. Three separate native signer volunteers have helped capture the Russian Signs, each with five repeats of each sign. One female volunteer, age 25, and two male volunteers, ages 27 and 33, participated as volunteers for the IIWS sign gesture database collection. The data were separated into categories based on the words, type of movement, and hand positions. The data was then categorized into idle, beginning sign, and ending. The database contains approximately 800 handshapes for static sign gestures with seven different signers in 12595 clips, each lasting about $3sec-5sec$, yielding 377850 frame images.

B. PERFORMANCE EVALUATION METRICS

We assess all techniques using recognition accuracy criteria.

The performance evaluation of the proposed hDNN-SLR framework was carried out in three phases. In the first phase the proposed framework was evaluated by training through multiple epochs and analyzed using the sign gesture recognition accuracy. In the second phase of performance evaluation, the model was trained in three level, each level separately, and the signer-based performance of the proposed model was evaluated. For the first signer-based performance evaluation level, the network was trained with a single signer dataset and tested for unseen multi-signer dataset. For the second signer-based performance evaluation level, the network was trained with a three different signer dataset and tested for unseen multi-signer dataset. For the third signer-based performance evaluation level, the network was trained with a five signer dataset and tested for unseen multi-signer dataset. In the third phase the performance proposed framework was compared with other baseline models concerning the accuracy metrics using the WLASL database and thereby the scalability of the proposed model was evaluated.

C. EXPERIMENTAL RESULTS FOR THE PROPOSED ISLR FRAMEWORK

Our proposed hDNN-SLR has been implemented using the newly created ISL and RSL datasets. The implementation has been carried out using RTX 3060 GPU. The Network was trained for 140 epochs. The Categorical Cross-Entropy Loss was employed for estimating the loss. Cross-entropy is used to fine-tune the BiLSTM for successful feature learning and recognition, while L1-loss is smoothed using SGD as an optimizer. Early stopping is used to stop training to improve the model's accuracy. The experimental findings were plotted in a graph. The accuracy generated from training and validation for the ISL and RSL Dataset is illustrated in the graph provided in Figure 4. The accuracy of the validation dataset was 99.87 per cent, while the training accuracy was 98.71 per cent. The loss plot generated from the train and validation of the Static ISL and RSL Dataset is plotted and illustrated in Figure 5.

D. EXPERIMENTAL RESULT FOR ISLR WITH VARYING CONSTRAINTS

The proposed framework experimented with varying constraints for evaluating its performance. The performance record based on the absence of each SCD module is shown in Figure 6. The graphical plot illustrates the performance of the proposed recognition framework without extracting certain

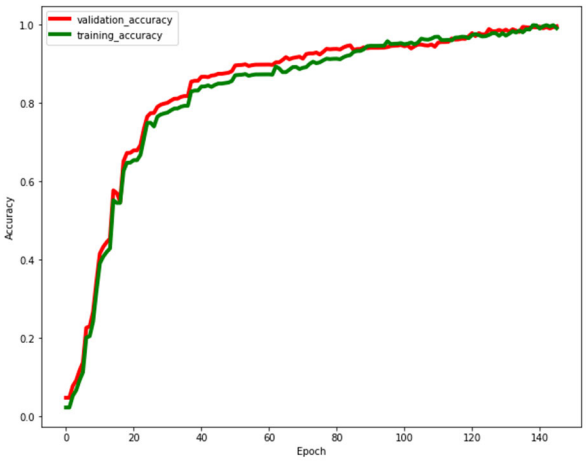

FIGURE 4. Accuracy plot for proposed hDNN-SLR.

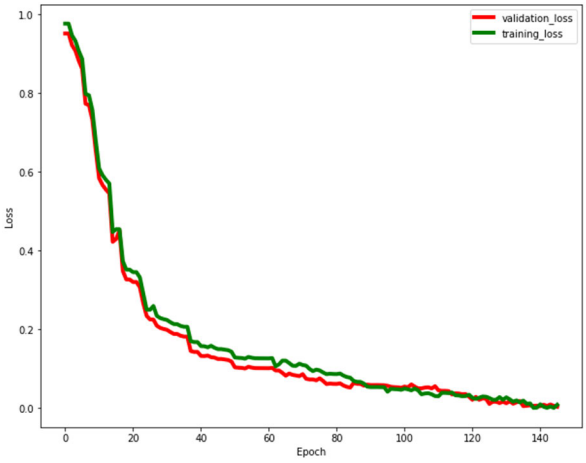

FIGURE 5. Loss plot for proposed hDNN-SLR.

features, including feature learning from Full-frame RGB images, posed extracted images, and Optical flow images. It also shows poor recognition accuracy when the non-manual features and the pose coordinates are not included during the feature learning process. When we consider all the modules from SCD, the recognition accuracy is noted to outperform by a large margin.

For evaluating the performance of the proposed framework with the multi-signer dataset, the test set has been divided concerning the number of signers using three constraints. The first constraint was to train the framework with a single signer train dataset and test the framework with a new signer validation dataset. Figure 7 illustrates the recognition performance for the first constraint. The second constraint was to train the framework with three different signer train datasets and test the framework with a set of new unseen signer validation datasets. Figure 8 illustrates the recognition performance for the second constraint. The third constraint was to train the framework with five different signer train datasets and test the framework with a set of new unseen signer validation datasets. Figure 9 illustrates the recognition performance for the third constraint. When comparing the three graphs with different multi-signer train-test ratios, we can see that the framework performs well and tends to attain convergence when the Network is well-trained with a multi-signer dataset.

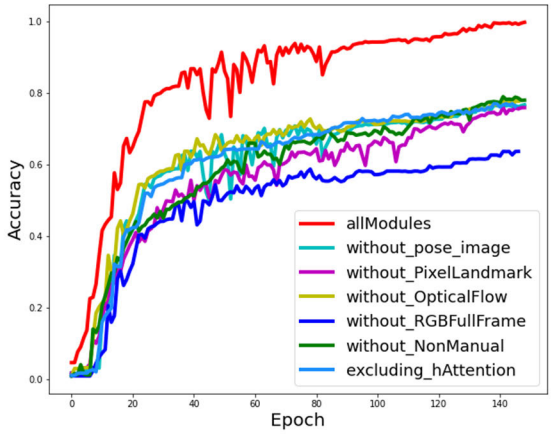

FIGURE 6. Accuracy concerning the presence of modules.

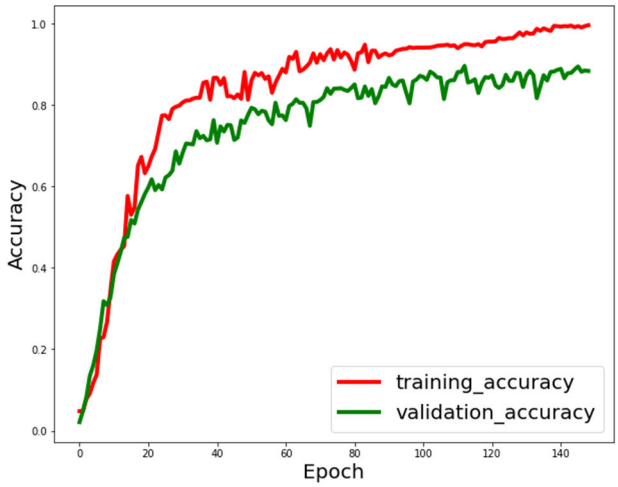

FIGURE 7. Performance of framework having trained with single signer and validation with the multi-signer dataset.

For evaluating the performance of the proposed framework with the multi-signer dataset, the test set has been divided concerning the number of signers using three constraints. The first constraint was to train the framework with a single signer train dataset and test the framework with a new signer validation dataset. Figure 7 illustrates the recognition performance for the first constraint. The second constraint was to train the framework with three different signer train datasets and test the framework with a set of new unseen signer validation datasets. Figure 8 illustrates the recognition performance for the second constraint. The third constraint was to train the framework with five different signer train datasets and test the framework with a set of new unseen signer validation datasets. Figure 9 illustrates the recognition performance for

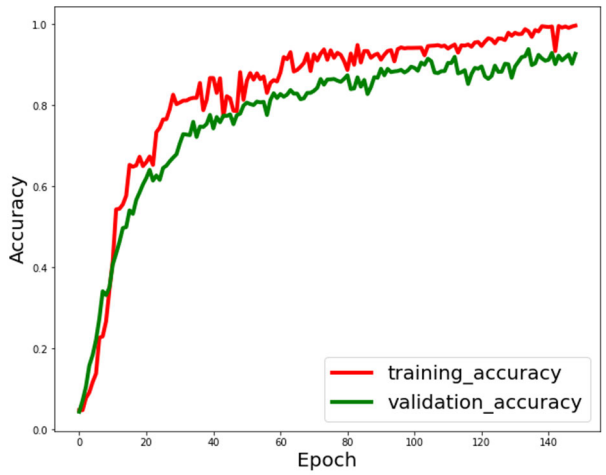

FIGURE 8. Performance of framework having trained with the multi-signer dataset with three different signer datasets and testing with the multi-signer dataset.

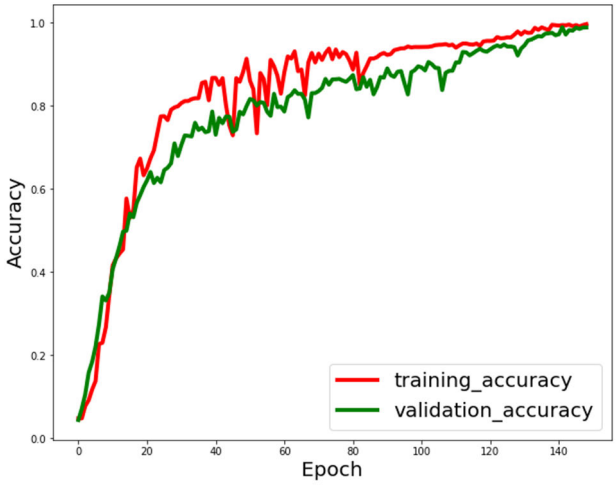

FIGURE 9. Performance of framework having trained with the multi-signer dataset with five different signer datasets and validation with the multi-signer dataset.

the third constraint. When comparing the three graphs with different multi-signer train-test ratios, we can see that the framework performs well and tends to attain convergence when the Network is well-trained with a multi-signer dataset.

E. EXPERIMENTAL RESULT FOR ISLR USING A BENCHMARK DATASET

The framework is implemented on WLASL Dataset, an American Isolated Word SL dataset. It comprises a collection of WLASL100, WLASL300, WLASL1000, and WLASL2000. The overall dataset's performance of the Top-1, Top-2, and Top-3 percentages was analyzed in terms of accuracy for each dataset group. The performance of the hDNN-SLR was evaluated with the other cutting-edge models such as I3D [43], Pose-TGCN [43], Pose-GRU [43], GCN-BERT [26], Multi-Stream [44], and Fusion-3 [45]. Table 2 shows the performance evaluation of hDNN-SLR with other baseline architectures in terms of Accuracy measures using the WLASL100, WLASL300, WLASL1000 and WLASL2000 dataset consisting of 100, 300, 1000 and 2000 word sign videos. Our proposed work tends to produce greater accuracy outcomes when compared to other baseline models.

TABLE 2. Performance accuracy (%) comparison of hDNN-SLR with other baseline architectures using benchmark datasets.

Model	WLASL-100	WLASL-300	WLASL-1000	WLASL-2000
I3D	89.92	86.98	84.33	66.31
Pose-GRU	85.66	76.05	70.15	61.38
Pose-TGCN	87.60	79.64	71.91	62.24
GCN-BERT	88.67	80.93	-	-
Fusion-3	90.16	86.22	84.71	75.71
Multi-Stream	96.05	94.83	92.94	87.47
Proposed hDNN-SLR	**98.75**	**98.02**	**97.94**	**97.54**

Nevertheless, when the quantity of the dataset grows, our proposed method appears to be more efficient and adaptable to a large vocabulary. The hDNN-SLR system we propose is scalable and adaptable for Multilingual SLR. The proposed work is better than other baseline models since it deals with multi-semantic feature learning rather than just considering only the manual feature. The proposed model considers all the constraints the signers face in real-life scenarios.

V. CONCLUSION AND FUTURE WORK

A novel Sign Language Recognition system has been developed by deploying a multi-semantic discriminative feature learning Deep Neural Net and spatial, temporal and sequential feature learning. The proposed hDNN-SLR framework intends to extract semantic manual co-articulations and non-manual elements, which are the key components necessary for sign recognition. In addition, spatial, sequential, and temporal features are also considered for accurate recognition. Furthermore, abstract and discriminative feature extraction is also carried out to segregate genuine and non-useful gestures. The genuine gestures are then used for recognizing the Sign gesture representations, thereby reducing the computation overhead. The experimentation of the proposed hDNN-SLR framework has been deployed on the newly created Indian and Russian Sign Language Datasets. The results generated represent a good and efficient performance. The performance of hDNN-SLR has also been compared with other baseline frameworks using the WLASL dataset. According to the analytical outcomes, the proposed hDNN-SLR paradigm surpasses the existing baseline architectures. Since the proposed work focuses on multi-semantic feature learning rather than merely taking into account the manual feature, it is superior to previous baseline models. The proposed model considers every limitation that signers encounter

in real-world situations. As a part of future work, we would like to extend our study toward continuous sign sentence recognition. We would also like to design a framework for handling the segmentation ambiguities and moment epenthesis in continuous sign language recognition. The isolated word gesture recognition framework could also be integrated to enhance sign spotting from continuous sign video stream for recognition sign sentences from continuous sign gestures. We also intend to increase the dataset and publicly publish it for further research.

ACKNOWLEDGMENT

We gratefully acknowledge the Department of Science & Technology (DST), India, for sanctioning the Indo-Russian Joint Project (INT/RUS/RFBR/393). We also acknowledge SASTRA Deemed University, Thanjavur, India, for extending infrastructural support to carry out this research.

REFERENCES

[1] Y. Saleh and G. F. Issa, "Arabic sign language recognition through deep neural networks fine-tuning," *Int. J. Online Biomed. Eng.*, vol. 16, no. 5, pp. 71–83, 2020.
[2] X. Jiang, M. Lu, and S.-H. Wang, "An eight-layer convolutional neural network with stochastic pooling, batch normalization and dropout for fingerspelling recognition of Chinese sign language," *Multimedia Tools Appl.*, vol. 79, nos. 21–22, pp. 15697–15715, Jun. 2020.
[3] O. Sevli and N. Kemaloğlu, "Turkish sign language digits classification with CNN using different optimizers," *Int. Adv. Researches Eng. J.*, vol. 4, no. 3, pp. 200–207, Dec. 2020.
[4] K. Wangchuk, K. Wangchuk, and P. Riyamongkol, "Bhutanese sign language hand-shaped alphabets and digits detection and recognition," Ph.D. dissertation, Dept. Comput. Eng., Naresuan Univ., Phitsanulok, Thailand, 2020.
[5] R. Elakkiya and E. Rajalakshmi. *ISLAN*. Mendeley Data. Accessed: Jan. 8, 2021. [Online]. Available: https://data.mendeley.com/datasets/rc349j45m5/1
[6] W. Sandler, "The phonological organization of sign languages," *Lang. Linguistics Compass*, vol. 6, no. 3, pp. 162–182, Mar. 2012.
[7] R. Elakkiya, "Retraction note to: Machine learning based sign language recognition: A review and its research frontier," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 7, pp. 7205–7224, Jul. 2022.
[8] S. Diwakar and A. Basu, "A multilingual multimedia Indian sign language dictionary tool," in *Proc. IJCNLP*, 2008, p. 57.
[9] S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," *Sign Lang. Stud.*, vol. 1064, no. 1, pp. 195–277, 1989.
[10] P. Eccarius and D. Brentari, "Symmetry and dominance: A cross-linguistic study of signs and classifier constructions," *Lingua*, vol. 117, no. 7, pp. 1169–1201, Jul. 2007.
[11] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using Kinect sensor," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1110–1120, Aug. 2013.
[12] C. Wang, Z. Liu, and S.-C. Chan, "Superpixel-based hand gesture recognition with Kinect depth camera," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 29–39, Jan. 2015.
[13] Z. Liu, X. Chai, Z. Liu, and X. Chen, "Continuous gesture recognition with hand-oriented spatiotemporal feature," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3056–3064.
[14] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4207–4215.
[15] R. Agarwal, "Bayesian K-nearest neighbour based redundancy removal and hand gesture recognition in isolated Indian sign language without materials support," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1116, no. 1, 2021, Art. no. 012126.
[16] O. Aran, "SignTutor: An interactive system for sign language tutoring," *IEEE MultiMedia* vol. 16, no. 1, pp. 81–93, Mar. 2009.
[17] I. N. Sandjaja and N. Marcos, "Sign language number recognition," in *Proc. 5th Int. Joint Conf. INC, IMS IDC*, 2009, pp. 1503–1508.
[18] Q. Yang, "Chinese sign language recognition based on video sequence appearance modeling," in *Proc. 5th IEEE Conf. Ind. Electron. Appl.*, Jun. 2010, pp. 1537–1542.
[19] S. Hore, "Indian sign language recognition using optimized neural networks," in *Information Technology and Intelligent Transportation Systems*. Cham, Switzerland: Springer, 2017, pp. 553–563.
[20] P. Malhotra, "Indian sign language recognition system using openpose," *I-Manager's J. Comput. Sci.*, vol. 7, no. 2, p. 43, 2019.
[21] S. Reshna and M. Jayaraju, "Spotting and recognition of hand gesture for Indian sign language recognition system with skin segmentation and SVM," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2017, pp. 386–390.
[22] J. Raheja, A. Mishra, and A. Chaudhary, "Indian sign language recognition using SVM," *Pattern Recognit. Image Anal.*, vol. 26, no. 2, pp. 434–441, 2016.
[23] K. Shenoy, T. Dastane, V. Rao, and D. Vyavaharkar, "Real-time Indian sign language (ISL) recognition," in *Proc. 9th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2018, pp. 1–9.
[24] A. Tyagi and S. Bansal, "Feature extraction technique for vision-based Indian sign language recognition system: A review," in *Computational Methods and Data Engineering*. Singapore: Springer, 2021, pp. 39–53.
[25] M. Mukushev, "Evaluation of manual and non-manual components for sign language recognition," in *Proc. 12th Lang. Resour. Eval. Conf., Eur. Lang. Resour. Assoc. (ELRA)*, 2020, pp. 1–6.
[26] A. Tunga, S. V. Nuthalapati, and J. Wachs, "Pose-based sign language recognition using GCN and BERT," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2021, pp. 31–40.
[27] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7784–7793.
[28] 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.
[29] S. Aly and W. Aly, "DeepArSLR: A novel signer-independent deep learning framework for isolated Arabic sign language gestures recognition," *IEEE Access*, vol. 8, pp. 83199–83212, 2020.
[30] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring cross-domain knowledge for video sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6205–6214.
[31] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 22965–22987, Aug. 2020.
[32] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3413–3423.
[33] S. Sharma and S. Singh, "Recognition of Indian sign language (ISL) using deep learning model," *Wireless Pers. Commun.*, vol. 123, no. 1, pp. 671–692, Mar. 2022.
[34] G. Jianchun, G. Jiannuan, and W. Lili, "Gesture recognition method based on attention mechanism for complex background," *J. Phys., Conf. Ser.*, vol. 1873, no. 1, Apr. 2021, Art. no. 012009.
[35] O. Mazhar, S. Ramdani, and A. Cherubini, "A deep learning framework for recognizing both static and dynamic gestures," *Sensors*, vol. 21, no. 6, p. 2227, Mar. 2021.
[36] M. Punchimudiyanse and R. G. N. Meegama, "Animation of fingerspelled words and number signs of the Sinhala sign language," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 4, pp. 1–26, Dec. 2017.
[37] A. A. I. Sidig, H. Luqman, S. Mahmoud, and M. Mohandes, "KArSL: Arabic sign language database," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, Jan. 2021.
[38] J. Singha and K. Das, "Recognition of Indian sign language in live video," 2013, *arXiv:1306.1301*.
[39] P. Kumar and S. Kaur, "Sign language generation system based on Indian sign language grammar," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 4, pp. 1–26, 2020.
[40] H.-Y. Jung, J.-H. Lee, E. Min, and S.-H. Na, "Word reordering for translation into Korean sign language using syntactically-guided classification," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 2, pp. 1–20, Mar. 2020.
[41] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[42] M. E. R. Grif and A. B. M. R. E. Prikhodko, "Recognition of Russian and Indian sign languages based on machine learning," *Anal. Data Process. Syst.*, vol. 3, no. 83, pp. 53–74, 2021.

[43] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1459–1469.
[44] M. Maruyama, S. Ghose, K. Inoue, P. P. Roy, M. Iwamura, and M. Yoshioka, "Word-level sign language recognition with multi-stream neural networks focusing on local regions," 2021, *arXiv:2106.15989*.
[45] A. A. Hosain, P. Selvam Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka, "Hand pose guided 3D pooling for word-level sign language recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3429–3439.
[46] Y. Hou, Q. Li, C. Zhang, G. Lu, Z. Ye, Y. Chen, L. Wang, and D. Cao, "The state-of-the-art review on applications of intrusive sensing, image processing techniques, and machine learning methods in pavement monitoring and analysis," *Engineering*, vol. 7, no. 6, pp. 845–856, Jun. 2021.
[47] Y. Xu, M. Kong, W. Xie, R. Duan, Z. Fang, Y. Lin, Q. Zhu, S. Tang, F. Wu, and Y.-F. Yao, "Deep sequential feature learning in clinical image classification of infectious keratitis," *Engineering*, vol. 7, no. 7, pp. 1002–1010, Jul. 2021.
[48] E. Rajalakshmi, R. Elakkiya, A. L. Prikhodko, M. G. Grif, M. A. Bakaev, J. R. Saini, K. Kotecha, and V. Subramaniyaswamy, "Static and dynamic isolated Indian and Russian sign language recognition with spatial and temporal feature detection using hybrid neural network," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 1, pp. 1–23, Jan. 2023.
[49] C. Zach, T. Pock, and H. Bischof, "A duality-based approach for real-time TV-L^1 optical flow," in *Pattern Recognition*. Berlin, Germany: Springer, 2007, pp. 214–223.

E. RAJALAKSHMI received the B.E. degree in information technology from the Cummins College of Engineering, Pune, in 2018, and the M.Tech. degree in computer science and engineering from SASTRA Deemed University, Thanjavur, in 2020, where she is currently pursuing the Ph.D. degree.

She is also working as a Project Associate with SASTRA Deemed University. She has contributed various papers and chapters for many high-quality Scopus and SCI/SCIE-indexed journals, conferences, and books. Her research interests include sign language recognition, music emotion recognition, deep neural networks, image processing, and computer vision. She is a Lifetime Member of the International Association of Engineers and a member of the Association for Computing Machinery.

R. ELAKKIYA received the Ph.D. degree from Anna University, Chennai, in 2018. She is currently working as an Assistant Professor with the Department of Computer Science, Birla Institute of Technology and Science, Pilani, Dubai Campus, Dubai International Academic City Dubai, United Arab Emirates. She has also worked as an Assistant Professor with the Department of Computer Science and Engineering, School of Computing, SASTRA University, Thanjavur. She got three patents and has published more than 35 research papers in leading journals, conference proceedings, and books, including IEEE, Elsevier, and Springer. Her research interests include deep learning and computer vision. She is a Lifetime Member of the International Association of Engineers. She is also an Editor of the *Information Engineering and Applied Computing* journal. She has organized various events, including a Workshop on "Cyber Security" at the Agni College of Technology, Chennai, from 2014 to 2017, sponsored by IIT Bombay, and a Software Development Workshop, "EKTIA" at the Jerusalem College of Engineering, Chennai, in 2012.

V. SUBRAMANIYASWAMY is currently working as a Professor with the School of Computing, SASTRA Deemed University, Thanjavur, Tamil Nadu. His research interests include machine learning, cognitive science, the Internet of Things (IoT), recommender systems, social network data mining, and big data analytics. His research aims to respond to the new trend and huge need for big data and cognitive computing specialists by filling a gap in multidisciplinary fields, including computer science, mathematics, engineering, and application. He has organized several national/international seminars/workshops/symposiums/conferences/special sessions in machine learning, cognitive computing, and big data. With an academic experience of more than 16 years, he serves as a guest editor and a reviewer for several IEEE, Springer, and Elsevier journals.

L. PRIKHODKO ALEXEY received the master's degree from the Department of Automated Control Systems, Novosibirsk State Technical University (NSTU). He is currently working as a Junior Researcher at the Department of Automated Control Systems, Faculty of Automation and Computer Engineering, NSTU. His research interests include deep learning, gesture recognition, and natural language processing. He has published 15 scientific papers. He won a DAAD Scholarship for an internship in Germany, in 2018.

GRIF MIKHAIL is currently working as a Professor at the Department of Automated Control Systems, Faculty of Automation and Computer Engineering, Novosibirsk State Technical University. In 1987, he defended his candidate's dissertation, and his doctoral dissertation, in 2002. He has published more than 300 scientific tutorials. His research interests include designing complex systems and computer sign language translation systems. For many years, he is the scientific director of grants and contractual research and development on decision-making systems in various industries.

MAXIM BAKAEV received the master's degree in digital design from Kyungsung University, South Korea, and the Ph.D. degree in software engineering, in 2012. He currently works as an Associate Professor with the Automated Control Systems Department, Novosibirsk State Technical University (NSTU), Russia. His research interests include human–computer interaction, universal design, web user interfaces, user behavior models, knowledge engineering, and machine learning. He has served as a committee member for several international conferences, particularly as the PC Co-Chair for ICMSC 2018 and ICWE 2019, the Demo Posters Chair for ICWE 2020, and the Workshops Co-Chair for ICWE 2021. He was a Reviewer for several international conferences and journals, including CHI, the *International Journal of Human-Computer Studies*, *Applied Ontology*, and *Symmetry*. In 2016, he received the Novosibirsk City Hall Award in science and innovations for Best Young Researcher in higher education institutions. Under his supervision, more than 20 bachelor's and master's students graduated.

KETAN KOTECHA is currently an Administrator and a Teacher of deep learning. He has expertise and experience in cutting-edge research and projects in AI and deep learning for the last 25 years. He has published over 100 widely in several excellent peer-reviewed journals on various topics ranging from cutting-edge AI, education policies, teaching-learning practices, and AI for all. His research interests include artificial intelligence, computer algorithms, machine learning, and deep learning. He was a recipient of the two SPARC projects worth INR 166 lakhs from the MHRD Government of India in AI in collaboration with Arizona State University, USA, and The University of Queensland, Australia. He was also a recipient of numerous prestigious awards, such as the Erasmus+ Faculty Mobility Grant to Poland, the DUO-India Professors Fellowship for research in responsible AI in collaboration with Brunel University, U.K., the LEAP Grant at Cambridge University, U.K., the UKIERI Grant with Aston University, U.K., and a Grant from the Royal Academy of Engineering, U.K., under Newton Bhabha Fund. He has published three patents and delivered keynote speeches at various national and international forums, including at the Machine Intelligence Laboratory, USA, IIT Bombay, under the World Bank Project, the International Indian Science Festival organized by the Department of Science and Technology, Government of India, and many more. He is an Associate Editor of IEEE Access journal.

LUBNA ABDELKAREIM GABRALLA received the B.Sc. and M.Sc. degrees in computer science from the University of Khartoum, and the Ph.D. degree in computer science from the Sudan University of Science and Technology, Khartoum, Sudan. She became a Senior Fellow (SFHEA), in 2021. She is currently an Associate Professor at the Department of Computer Science and Information Technology, Princess Nourah Bint Abdulrahman University, Saudi Arabia. Her research interests include soft computing, machine learning, and deep learning.

AJITH ABRAHAM (Senior Member, IEEE) received the M.Sc. degree from Nanyang Technological University, Singapore, in 1998, and the Ph.D. degree in computer science from Monash University, Melbourne, Australia, in 2001. He is currently the Director of the Machine Intelligence Research Labs (MIR Labs), a Not-for-Profit Scientific Network for Innovation and Research Excellence Connecting Industry and Academia. With its HQ in Seattle, USA, the Network currently has over 1,500 scientific members from over 105 countries. As an Investigator/a Co-Investigator, he has won research grants worth more than U.S. 100 million. He also holds two university professorial appointments. He works as a Professor in artificial intelligence at Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair of Artificial Intelligence at UCSI, Malaysia. He works in a multidisciplinary environment. He has authored/coauthored more than 1,400 research publications, of which there are more than 100 books covering various aspects of computer science. One of his books was translated into Japanese, and a few other articles were translated into Russian and Chinese. He has more than 46 000 academic citations (H-index of more than 102 as Per Google Scholar). He has given over 150 plenary lectures and conference tutorials (in more than 20 countries). He was the Chair of the IEEE Systems Man and Cybernetics Society and the Technical Committee on Soft Computing (with more than 200 members), from 2008 to 2021. He served as a Distinguished Lecturer for the IEEE Computer Society representing Europe, from 2011 to 2013. He was the Editor-in-Chief of *Engineering Applications of Artificial Intelligence* (EAAI), from 2016 to 2021, and served on the editorial board for over 15 international journals indexed by Thomson ISI.