SIGNGRAPH: A faster and efficient Graph Convolution method for Pose-based Sign Language Recognition

Neelma Naz[1], Hasan Sajid[1], Sara Ali[1], Osman Hasan[1], and Muhammad Khurram Ehsan[2]
[1]National University of Sciences and Technology (NUST), Islamabad, 44000, Pakistan
[2]Faculty of Engineering Sciences, Bahria University Lahore Campus, Lahore 54600, Pakistan

Corresponding author: Neelma Naz (e-mail: neelma.naz@seecs.edu.pk).

ABSTRACT Sign language recognition (SLR) enables the deaf and speech-impaired community to integrate and communicate effectively with the rest of society. Word level or isolated SLR is a fundamental yet complex task with the main objective of using models to correctly recognize signed words. Sign language consists of very fast and complex hand, body, face movements, and mouthing cues that make the task very challenging. Several input modalities; RGB, optical Flow, RGB-D, and pose/skeleton have been proposed for SLR. However, the complexity of these modalities and the state-of-the-art (SOTA) methodologies tend to be exceedingly sophisticated and over-parametrized. In this paper, our focus is to use the hands and body poses as an input modality. One major problem in pose-based SLR is extracting the most valuable and distinctive features for all skeleton joints. In this regard, we propose an accurate, efficient, and lightweight model based on a graph convolution network (GCN) along with residual connections and a bottleneck structure. The proposed model not only facilitates efficient learning during model training providing SOTA accuracies but also alleviates computational complexity. With the proposed model in place, we are able to achieve SOTA accuracies on three different subsets of the WLASL dataset and the LSA-64 dataset. Our proposed model outperforms previous SOTA pose-based methods by providing a relative improvement of 8.91%, 27.62%, and 26.97% for WLASL-100, WLASL-300, and WLASL-1000 subsets. Moreover, our proposed model also outperforms previous SOTA appearance-based methods by providing a relative improvement of 2.65% and 5.15% for WLASL-300 and WLASL-1000 subsets. For the LSA-64 dataset, our model is able to achieve 100% test recognition accuracy. We are able to achieve this improved performance with far less computational cost as compared to existing appearance-based methods.

INDEX TERMS Graph Convolution Network (GCN), Skeleton Modeling, Sign Language Recognition

I. INTRODUCTION

According to the World Federation of the Deaf, there are over 70 million people with hearing and speech impairments and over 300 sign languages are used by these people [1]. This means that a large portion of the world population suffers from this communication barrier that directly affects their daily interactions and causes inequality in society. Sign language (SL) is a primary communication tool for deaf people. Therefore, the design of efficient mechanisms for automatic sign language recognition (SLR) will not only break down this communication barrier but will increase the availability of opportunities for a major portion of the world's population. In contrast to written and spoken languages, sign languages make use of "corporal-visual" channels produced by body motions and interpreted by eyes. This makes automatic SLR a very interesting research domain that requires expertise from computer vision as well as natural language processing to efficiently understand the spatio-temporal linguistic constructs of performed signs. Just like other languages, sign languages have their underlying structures, grammar, inter alia, intricacies, and articulators that allow users to effectively communicate and express themselves. Multiple channels/articulators are used by signers to convey complex semantics [2]. These channels can be categorized into two main groups, i.e., manual or non-manual features [3] based on their role in information communication. Manual features represent the macro motions, like the hand and arm movements as well as hands shape, location, and palm orientation. Although manual features play a dominant part in sign morphology, they cannot encapsulate the full spatial and temporal context of the information being conveyed. To fill this gap and provide additional information, clarity, and context, non-manual features, including body pose, facial expressions, and mouthing cues, are used. Manual and non-manual features

are often used together and thus affect each other's meanings. Many applications such as interpreting services, translation systems, human-computer interactions [4], real-time person recognition systems, virtual reality, robot controls, games [5] and hand tracking in desktop environments [6] can benefit from advancements in SLR.

SLR can be categorized into two domains: Isolated/Word level SLR and Continuous SLR. Isolated SLR recognizes or classifies individual sign recordings into correct glosses (signed words) while Continuous SLR translates the whole utterances containing multiple glosses. A special case of isolated SLR is the classification of gestures into alphabets.

SLR techniques make use of videos or data acquired through wearable sensors mounted on a hand glove as input. One of the early approaches to hand gesture recognition dates back to 1987 where magnetic flux sensors mounted on a glove were used to acquire hand position and orientation [7]. With the latest advancements in deep learning, visual sign language recognition has gained much attraction. However, visual recognition faces many challenges (i.e. occlusions, illumination changes, different viewpoints, different image resolutions, and cluttered backgrounds) making it more complex to design a universal automatic SLR model.

From a computer vision perspective, visual sign language recognition mainly depends upon visual features acquired through RGB or RGB+D cameras. These features are then used for extracting strong and efficient spatiotemporal representations that encapsulate visual characteristics, such as hand shape, palm orientation as well as motion profile of hands and arms. In recent years, several deep learning based approaches are proposed that can learn the most useful spatiotemporal relations in videos and are being used for action recognition, gait recognition, and action localization [8-10]. The latest deep learning models make use of ConvNets to extract spatial cues and Recurrent neural networks to model temporal dependencies. Some of these models make use of 3D ConvNets [11] to fuse spatial and temporal cues. However, appearance-based methods have distinctively higher computational complexity originating from higher data dimensionality. Some other methods make use of human hands and body poses extracted by efficient pose estimation algorithms [12-14]. Human poses are represented by the locations of body and hand joints, bones, and facial landmarks [15-17]. In general, models using skeleton data as input are lightweight, compute-efficient, and suitable for edge devices that dramatically increase their potential in everyday use. Despite these advantages, pose-based methods for SLR still suffer from low accuracies.

Motivated by the substantially low accuracies and high computational efficiency of pose-based methods, we have proposed a highly accurate, lightweight method for isolated SLR using spatiotemporal graph convolutions and residual mechanisms. We have shown that our proposed model provides SOTA results while being computationally efficient which makes it a perfect choice for an accurate and lightweight sign language recognition solution capable of running on edge devices.

In summary, the main contributions of our work include:

- An accurate and lightweight graph convolution Network with residual connections and bottleneck structure is proposed that is capable of effectively capturing spatiotemporal dependencies in the input poses.
- Constituting SOTA on the WLASL-100, WLASL-300, and WLASL-1000 datasets when compared with pose-based SLR methods with the performance improvement of 8.91%, 27.62%, and 26.97% on 100, 300 and 1000 glosses subsets, respectively.
- Constituting SOTA on the WLASL-300 and WLASL-1000 subsets when compared with appearance-based methods with performance improvement of 2.65% & 5.15%.
- Constituting SOTA for the LSA-64 dataset by achieving 100% test recognition accuracy.
- Constituting improved computational efficiency by 20× reduction in the number of model parameters, 2.68× reduction in FLOP computations, and 13.6× reduction in inference time.

II. RELATED WORKS

SLR has achieved substantial progress in recent years in terms of recognition accuracy. The emergence of deep learning-based architectures and the availability of high computational resources has made the design and implementation of deep models using multi-modal data a possibility. The problem of automatic SLR mainly involves three phases: choosing the appropriate input modality, extracting spatiotemporal features from the input data, and a prediction phase. All these phases have been approached in several ways. There are various data modalities, like RGB, RGB+D, and 2D or 3D skeleton features, that can be considered appropriate as input to the feature extraction modules. In the early days of SLR, hand-crafted features, like Histogram of Gradients (HOG) based features, Scale Invariant Feature Transform (SIFT) based features, motion velocity vector, and frequency domain features [18-21], were used to generate spatial representations. Whereas temporal dependencies were extracted, using condition random fields, Hidden Markov Models (HMM) [22, 23], and Dynamic Time Warping, to handle variable frame rates. However, these methods lack generalization ability. The final prediction phase was treated as a classification problem. To predict the sign classes, Support Vector Machine (SVM) were used.

The task of SLR shares the same problem structure as gesture and action recognition. So, the approaches to solve the SLR problem are mostly inspired by network architectures proposed for action and gesture recognition and can be classified into mainly three categories based on the input data modality:

- Appearance based Methods
- Pose/Skeleton based Methods

- Hybrid Methods

A. APPEARANCE BASED SLR WITH SEQUENTIAL/DYNAMIC INPUT DATA

Appearance-based methods mainly focus on spatial features in each frame, i.e., hand shapes, locations, orientations, and sometimes facial clues, and temporal features in a sequence of frames, i.e., hand, arms, and sometimes body motion. As the background is not useful in sign recognition, it is subtracted from the input image. To extract spatial features, 2D-convolutional neural network (CNN) based deep models [24, 25] are useful. Whereas to capture temporal information, Recurrent neural networks (RNNs) are used. Some studies make use of both traditional and deep learning based methods. In [26], hand-crafted features, like hand shapes and locations, are estimated using a single-shot detector (SSD) model and fed to a CNN model to extract spatial features, which are then fused and provided to a Long-Short term memory (LSTM) model for temporal feature extraction.

To learn spatiotemporal patterns in video frames, 3D-CNNs are also a popular choice and have shown remarkable performance. A large-scale dataset MS-ASL is proposed in [27] and a baseline is established using 2D- CNN followed by an LSTM module and 3D-CNN model. A 3D-CNN based I3D model baseline is proposed in [28]. Various other appearance-based baselines have also been proposed in [15] including a) 2D CNN + Gated Recurrent Unit (GRU) and b) 3D-CNN claiming the best results obtained by the I3D network. In [29], a SLR and education system is proposed. This SLR system is built upon a spatiotemporal network for semantic category identification of a given sign video while the education system detects the failure mode of learners and guides them to sign correctly. SLR is treated as a zero-shot learning problem in [30] to efficiently use the models learned on the seen sign glosses to unseen sign glosses. Textual sign descriptions and attributes, such as hand shapes, orientations, and right, left, or both hands used to perform a sign, are collected from sign language dictionaries. This information is then used as auxiliary data to learn semantic class representations.

The usage of RGB+D (Depth) images as input have also been studied in the latest literature. The depth stream assists in learning more complex features by ignoring the video background. Recently, the SUGO model based on 3D-CNN is proposed that uses data acquired through LIDAR [31]. Appearance-based methods for SLR critically suffer from high computational complexity in terms of memory requirements and processing power and have significantly low accuracies.

B. APPEARANCE-BASED SLR WITH MHI/STATIC INPUT DATA

Motion history image (MHI) is a static grey scale or colored image that represents the history of the motion present in a video images sequence into a single image. MHI is computed by various methods. Colored MHIs, namely star RGBs, are created in [32] to represent video sequences. These MHIs are fed to two ResNets and their features are combined using an attention mechanism. Three different types of motion templates: RGB motion image, dynamic image, and MHI have been created [33]. A single colored MHI representing the entire sequence is proposed in [34] and is combined with the I3D model to learn spatiotemporal dependencies in a sign video.

C. SKELETON-BASED METHODS

With the latest advancements in pose estimation methods, SLR based on pose data is receiving an increasing attention. Pose estimation involves the extraction of 2D or 3D skeletal joint data from an image or a video sequence. The methods to localize joints are divided into two main categories: Top-down [12] (Localize the human first and then localize body parts) or bottom-up approaches [14] (localize body parts and then group them). Skeleton-based SLR methods take body pose, hand pose, and sometimes face data as input. In [15], the body pose sequence is extracted and a gated recurrent unit (GRU) is employed to extract spatiotemporal clues. Additionally, a temporal Graph Convolutional Network (TGCN) is also employed on this pose data and a baseline is provided. In [16], spatial and temporal information has been captured separately using GCNs and a BERT model, and late fusion is performed to make the final predictions. A transformer-based model is employed in [17] for pose-based SLR. Although previously proposed pose-based SLR methods are less accurate as compared to appearance- based methods but are computationally cheaper and efficient.

D. HYBRID/MULTI-MODAL METHODS

Using multiple modalities, like RGB, depth, optical flow, and skeleton, is a common approach. These modalities are fused together by either early fusion or late fusion strategies to extract the most useful features to enhance recognition accuracies. The winning teams of the ChaLearn-2021 challenge used different types of data modalities [35, 36], e.g., skeleton, optical flow, RGB, depth, depth flow, and depth HHA, and used these multi-modal ensembles to improve the accuracy. In [26], several manual and non-manual features have been extracted from input videos. Firstly, several 2D-CNN-LSTM models have been trained separately using RGB, depth, and optical flow data and then these features are fused at the classification level using the best 2D-CNN-LSTM model. In [11], both motion and hand shape cues have been used as input features and fed to a 3D-CNN. A pose-guided 3D pooling mechanism is used to fuse the prediction score during test time. Although hybrid or multi-modal methods perform well in terms of accuracy but require much more processing power than methods using a single input modality.

E. SIGN LANGUAGE DATASETS

TABLE I
STATISTICS OF EXISTING SLR DATASETS

Datasets	Year	Language	Modalities	#Classes	#Signers	#Videos
WLASL [18]	2020	American	RGB	2000	119	21,803
AUTSL [22]	2020	Turkish	RGB+ Depth +skeleton	226	43	36,302
MS-ASL [19]	2019	American	RGB	1000	222	25,513
LSA64	2016	Argentine	RGB	64	10	3,200
DGS	2012	German	RGB + Depth	40	15	3,000

There are several publicly available sign language datasets that target different sign languages, word level or continuous sign languages, data sizes, number of signers, sensors to capture the data, and signer dependency. The most significant datasets reported in the literature include WLASL [15] - a large-scale American sign language dataset that has a vocabulary of 2000 glosses. MS-ASL [27] is also an American sign language dataset that includes a collection of publicly recorded videos of American sign language. LSA64- is an Argentinian sign language dataset with video recordings having colored hand gloves to make hand segmentation easier [37]. DGS Kinect 40 [38] is a German sign language dataset and includes RGB+D images. AUTSL [39] is the most recent Turkish sign language dataset. The details of these datasets are provided in Table I.

Our proposed model is evaluated on the LSA-64 and the Word Level American Sign Language (WLASL) dataset [15]. WLASL is the largest signer-independent American sign language dataset, collected from 20 different public websites and signs are performed by American signers or interpreters. The presence of a variety of dialects, signing styles, and video backgrounds makes it quite challenging. The authors of WLASL have provided four subsets of data named WLASL-100, WLASL-300, WLASL-1000, and WLASL-2000, where 100, 300,1000, and 2000 represent the number of glosses present in a subset. Details of LSA-64 and WLASL subsets are given in Table II.

TABLE II
DETAILS OF WLASL-SUB SPLITS AND LSA-64

WLASL-Subset	#Glosses	#Videos	#Signers
WLASL-100	100	2,038	97
WLASL-300	300	5,117	109
WLASL-1000	1000	13,168	116
WLASL-2000	2000	21,083	119
LSA-64	64	3200	10

III. PROPOSED METHODOLOGY

In this paper, we propose a skeleton/pose based SLR model for recognizing isolated signs. In this section, first a formal problem definition is provided, and then the main components of proposed approach are explained.

A. PROBLEM DEFINITION

Given a training dataset $\chi_{tr} = \{(x_i, c_i)\}_{i=1}^{N}$ that consists of N sample videos and $\{x_i \in \mathcal{R}^{T\times H\times W\times 3}\}$ is the i[th] training video and $c_i \in C_s$ the corresponding sign class; where H, W, 3 represent the height, width, and channel information of a single frame, respectively, and T is the total number of frames in the video. We aim to extract pose information of each video sample using the pose estimation method, such that $\{x_i \in \mathcal{R}^{T\times V\times C}\}$ where T, V, and C represent total frames, number of nodes, and node features and to construct a graph using this skeletal data. We aim to design an efficient and lightweight graph convolution network (GCN) based model to extract spatiotemporal dependencies of this sequential data to correctly predict the sign class labels.

B. POSE EXTRACTION

Historically, several methods have been reported to estimate human pose from RGB images or video sequences [12-14, 40, 41]. Most of these methods can measure human body pose efficiently but fail to correctly estimate hand joints and pose. SLR is highly dependent on hand shapes and poses, so the importance of an efficient pose estimator cannot be undermined. We have used MediaPipe Holistic [42], a multistage pipeline developed by MediaPipe, to preprocess the data and extract pose features from the image. Fig.1 clearly shows the overall functionality of MediaPipe Holistic model.

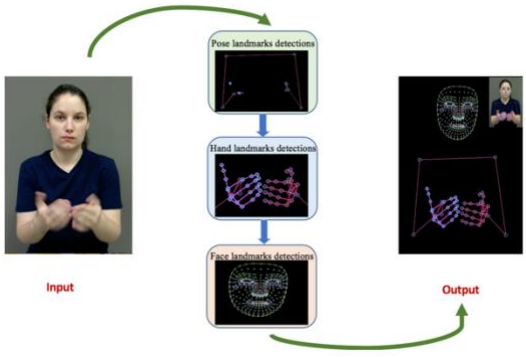

Figure 1. Overview of MediaPipe Holistic Pose Estimator

MediaPipe is an open-source framework that processes perceptual data, such as videos and images, by creating pipelines with a hybrid platform. For each frame in the input video, MediaPipe Holistic uses individual models i.e., MediaPipe holistic hand landmarks, MediaPipe holistic pose landmarks, and MediaPipe holistic face landmarks detector for pose estimation of the hands, body, and face regions. The MediaPipe holistic pose landmarks detector estimates 33 3D

landmarks from the given image or video frame consisting of x, y, and z coordinates. Whereas MediaPipe holistic hand pose model estimates 21 3D landmarks for each hand. It also provides binary classification of hands (left and right hand), and a hand flag showing the hand presence probability in the input image. The MediaPipe holistic face model estimates 468 3D face landmarks by using a single input camera image.

Pose Extraction is the stage 1 of our proposed method and is briefly discussed below:

- Blaze Pose's pose detector was used to estimate human pose and subsequent landmarks model. Then, three regions of interest (ROI) crops representing face and each hand (left and right) were derived using inferred pose landmarks, and then to improve the ROI, a new re-crop model was employed.
- The full-resolution input coordinates were cropped to these ROIs for task specific hand and face models, and corresponding landmarks were estimated.
- Finally, these estimated landmarks were joined together to produce pose information.

This model generated a total of 540+ landmarks, out of which we have used data for only 65 landmarks. These 65 landmarks consist of pose information for both hands, arms, body torso and some significant facial nodes like eyes, nose, ears, and lips. We have discarded all the remaining landmarks because they were providing no additional information in our model.

C. DATA PRE-PROCESSING AND CLEANING

In total, data of 65 joints (21 joints for each hand, 11 head joints and 12 joints for the body) has been considered to be given as input to our model. The head joints include the eyes, ears, and nose. From the estimated body joints, we only retain upper body joints that contain the neck, shoulders, elbows, and wrists. The hand joints contain four joints for each finger and one joint for the wrist. The data from each frame is concatenated and stored in a file. Although, MediaPipe provides information on 3D (x, y, and z) landmarks, but the model is not fully trained to predict depth (z-coordinates) accurately [42]. Thus, for our network, we have used only 2D (x & y coordinates) features and discarded z-coordinates. The extracted pose has a dimension of $\{x_i \in \mathcal{R}^{T\times 65\times 2}\}$.

D. DATA NORMALIZATION and AUGMENTATION

MediaPipe Holistic pose estimation method provides landmark coordinates normalized to [0, 1] by the image width and height respectively. These coordinates are then shifted by [-0.5; -0.5] to make sure that mean is zero and the standard deviation is 1. Although we tested different mean and standard deviation values, the reported ones gave the best results. The resultant normalized coordinates are then multiplied by 2 to maintain a fixed scale. We have augmented the data using two noise transform, that creates a cropped copy of the given sequence.

E. GRAPH PRELIMINARIES

Notation. A human skeleton graph can be constructed as a unidirectional spatiotemporal graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ on a pose sequence with T frames and N joints, where $\boldsymbol{\mathcal{V}}$ is the set of N nodes $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N\}$ representing the body and hand joints. Spatial and temporal connections can be represented by connecting adjacent joints in spatial dimension and joining all joints to themselves in the temporal dimension. $\boldsymbol{\mathcal{E}}$ is representing edge connections and is captured by an adjacency matrix $\boldsymbol{\mathcal{A}} \in \boldsymbol{\mathcal{R}}^{N\times N}$, where $\boldsymbol{\mathcal{A}}_{i,j} = \mathbf{1}$ if node $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ is connected by an edge otherwise $\boldsymbol{\mathcal{A}}_{i,j} = \mathbf{0}$.
Each joints has a feature set $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_{t,n} \in \boldsymbol{\mathcal{R}}^{C} \,|\, \mathbf{1} \le \boldsymbol{t} \le \boldsymbol{T}, \mathbf{1} \le \boldsymbol{n} \le \boldsymbol{N}\}$, where C is the total number of features in a single joint, T represents the total number of frames in a video sequence and N is the number of joints in each frame. $\boldsymbol{\mathcal{X}}$ is characterized as a feature tensor of dimension $\boldsymbol{\mathcal{R}}^{T\times N\times C}$. Thus, each sign sequence can be defined structurally by $\boldsymbol{\mathcal{A}}$ and feature-wise by $\boldsymbol{\mathcal{X}}$ such that $\boldsymbol{X}_t \in \boldsymbol{\mathcal{R}}^{N\times C}$ is a pose at time t. C-dimensional pose feature $\boldsymbol{X}$ is a tuple of 2D coordinates.

F. GRAPH CONVOLUTION NETWORK

Graph convolutions are an integral part of our proposed architecture. Given skeleton inputs, defined by adjacency matrix $\boldsymbol{\mathcal{A}}$ and features vector $\boldsymbol{X}$, graph convolutions can be applied using layer-wise update rule to the features at time t as given in Equation (1).

$$X_t^{(l+1)} = \sigma\left(D^{-1/2}(\boldsymbol{\mathcal{A}} + I)D^{-1/2}X_t^{(l)}\theta^{(l)}\right) \qquad (1)$$

Where $\boldsymbol{\mathcal{A}}$ is the adjacency matrix and represents intra-body connections and the identity matrix (I) represents self-loops, D is the diagonal degree matrix of $(A+I)$, $\theta^{(l)}$ denotes trainable weight matrix and $\sigma(.)$ represents an activation function. Intuitively, $D^{-1/2}(\boldsymbol{\mathcal{A}} + I)D^{-1/2}X_t^{(l)}$ can be explained as an approximate spatial mean feature aggregation of the messages being passed by the direct neighbors. These are called spatial graph convolutions (SCN). The graph's temporal convolutions (TCN) can also be implemented as a standard 2D convolution with the kernel size of $L \times 1$ along the temporal dimension and with a reception field of L to aggregate the contextual information embedded in adjacent frames. L is a hyperparameter that defines the length of temporal window. A basic block is constructed by both spatial and temporal convolutions, where each convolution is followed by a Batch Normalization layer and RELU activation layer. We adopt an extended variation of spatial graph convolution called Residual GCN or ResGCN as proposed in [43] for activity recognition. We modify the same concept to our use case of SLR.

G. NETWORK ARCHITECTURE

We use ResGCN [43] as baseline model for sign language recognition. In this model, we construct basic and bottleneck blocks by using ST-GCN block. Spatial graph convolutions and *2D* temporal convolutions are sequentially executed to learn spatial features in a single frame and features temporal

dependencies in video frames. The model consists of two types of blocks, i.e., standard basic ST-GCN block and bottleneck block. A basic block consists of a spatial block and a temporal block as presented in Fig.2. A standard graph convolution followed by batch normalization and RELU activation function is implemented in spatial block whereas temporal block consists of standard 2D temporal convolution followed by batch normalization and RELU activation.

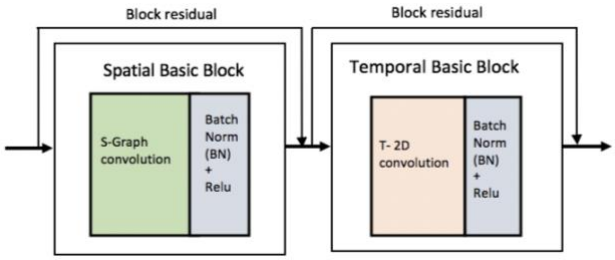

Figure 2. Basic Spatial and Temporal block structure

Inspired by ResNet [44], a bottleneck structure as presented in Fig.3 has been introduced in ResGCN. The bottleneck block consists of two *1×1* convolution layers with a reduction rate of *R* to reduce the number of feature channels. Each bottleneck convolution layer is added before and after the convolution layer. These bottleneck layers are added in both spatial and temporal blocks resulting in a reduction in the number of parameters. The original ST-GCN architecture has been further modified by the addition of residual connections. Residual connections were proposed in [43] to connect the features before and after every spatial and temporal block. Two types of residual connections have been proposed in this architecture, i.e., module residual, which connects the output of the previous basic or bottleneck block to the output of the current block and block residual, which connects the features before and after every temporal and spatial block.

The complete pipeline of proposed architecture is presented in Fig.4. Table III and IV give an overview of network's output after each stage.

IV. EXPERIMENTAL RESULTS

In this section, the performance of the proposed architecture is evaluated on three subsets of the WLASL dataset and LSA-64. Our results are compared with the SOTA appearance-based and pose based methods. Results of ablation studies are also provided to show the importance of various components in the proposed architecture.

TABLE III
OVERVIEW OF RESGCN-N21 ARCHITECTURE WITH REDUCTION RATE R=8

Module	Output
Batch Normalization	64× 65×2
Basic Block	64× 65× 64
Bottleneck Block1	64× 65× 32
Bottleneck Block2	32 × 65× 128
Bottleneck Block3	16 × 65× 256
AvgPool2D	1×256
FCN	1×embedding_size
FCN	1×num_classes

TABLE IV
OVERVIEW OF RESGCN-N39 ARCHITECTURE WITH REDUCTION RATE R=4

Module	Output
Batch Normalization	64× 65×2
Basic Block	64× 65× 64
Bottleneck Block1	64× 65× 64
Bottleneck Block2	64× 65× 32
Bottleneck Block3	32 × 65× 128
Bottleneck Block4	32 × 65× 128
Bottleneck Block5	16 × 65×256
Bottleneck Block6	16 × 65× 256
AvgPool2D	1×256
FCN	1×embedding_size
FCN	1×num_classes

A. DATASETS

We have tested the performance of the proposed architecture on the most recent American SL dataset: WLASL and Argentinian SL dataset LSA-64. We have performed experiments on three subsets of WLASL, WLASL-100, WLASL-300, and WLASL-1000. Each subset has its specifications as described in Table II. As the dataset is

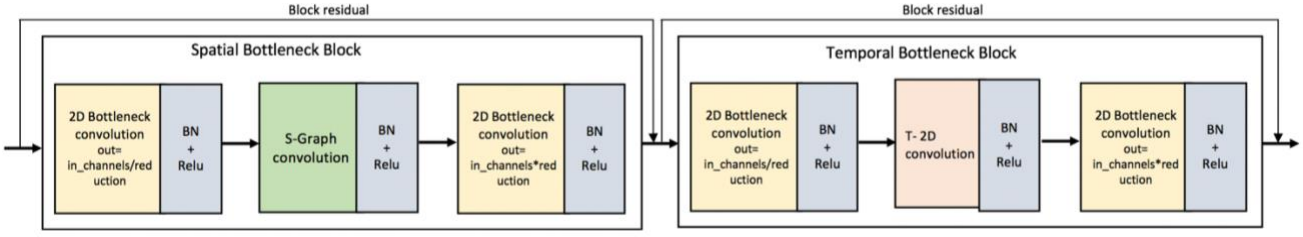

Figure 3. Bottleneck Spatial and Temporal Block structure along with block residual connections

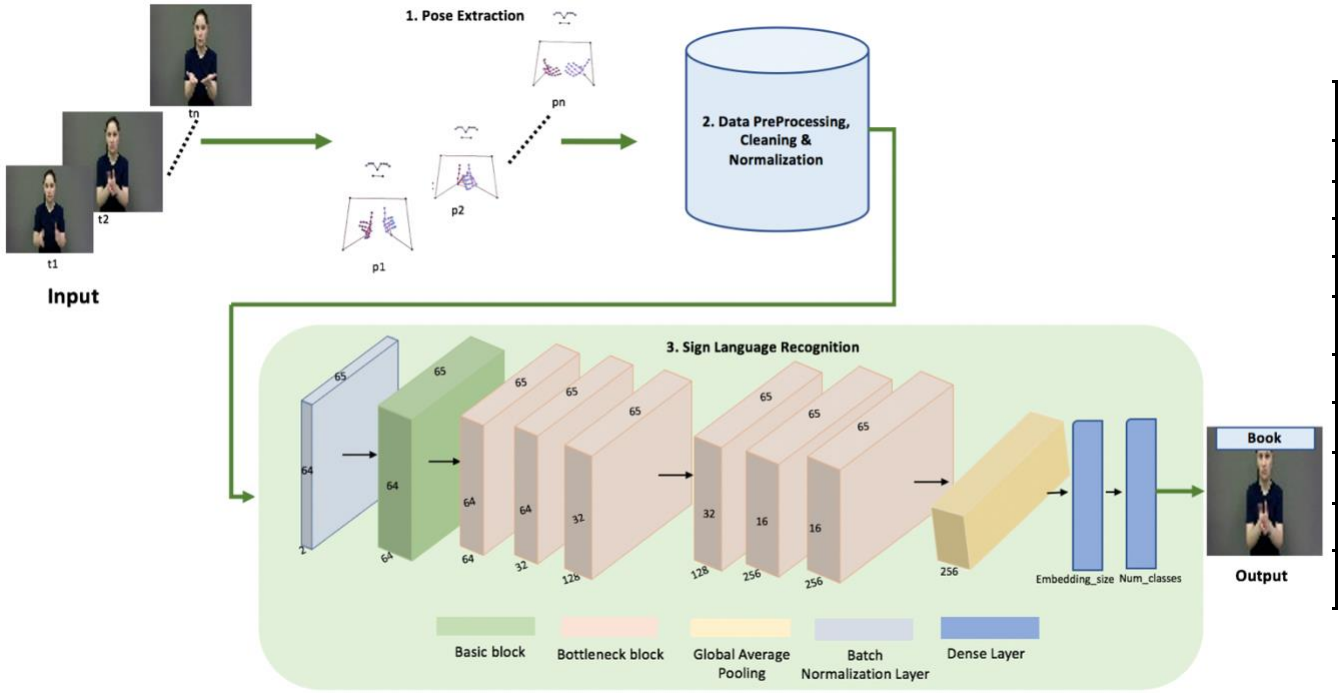

Figure 4. **Overall pipeline of proposed architecture**

collected from various online open-source resources so it has a huge diversity in terms of the number of signers, illumination conditions, and background variations making it a challenging dataset. Given the video sequences, body and hand poses are extracted using the MediaPipe pose extraction method. We have opted for the same training and testing protocols as used by dataset providers [15]. LSA-64 contains data of 3200 videos with 50 video samples per class signed by ten nonexpert signers. We have used 80:20 % split for training and testing purposes.

B. IMPLEMENTATION DETAILS

For implementation, we have partitioned the pose sequence as a graph using the spatial configuration provided in [8]. Experiments are carried out on a single NVIDIA 3080 RTX-GPU using PyTorch. Model is trained for 350 epochs. Random start sampling strategy is used to choose the required sequence length. For cases, where video lengths are shorter than required sequence length, frame sequence is appended with the last frame of the given video. The embedding size is chosen to be equal to number of classes. Further details of hyperparameters are given below.

- Optimizer: Adam optimizer [45]
- Temporal Kernel Size: 9
- Batch Size: 32
- Sequence Length: 64
- Graph Distance: 2
- Scheduler: cyclic learning rate schedular with learning rate 0f 0.01
- Loss Function: Cross Entropy

C. COMPARISON WITH STATE-OF-THE-ART

We report the top-1, top-5, and top-10 accuracy of the proposed architecture on the WLASL-100, WLASL-300, and WLASL-1000 datasets. We also compare our method with the SOTA as presented in Table V. For comparison purposes, we divide the methods into two main sections: the first section represents the appearance-based methods (methods taking RGB sequences as input) for SLR, whereas the second represents only the skeleton or pose-based models.

By using SIGNGRAPH, we are able to surpass the previous SOTA pose-based approach's accuracy by 8.91% for WLASL-100 subset, by 27.62% for WLASL-300 subset and by 26.97% for WLASL-1000 subset. Our model is also able to outperform SOTA appearance-based methods by 2.65% and 5.15% for WLASL-300 and WLASL-1000 subsets. For WLASL-100 subsets, our model produces comparable results with appearance-based methods. It is evident from our results, as number of glosses to be classified increases, our method is robust while all other methods fall apart.

Moreover, there are some ambiguities in the dataset, i.e., two different words signed in the same way or the same word signed differently by different users. These ambiguities make it very hard even for humans to differentiate between the glosses. These ambiguous behaviors mislead the deep learning models resulting in lower accuracies as can been seen in Fig.5. Last two rows of Fig.5 show that the same word "before" is signed differently by different signers resulting into incorrect gloss predictions. As the size of dataset increases, the number of ambiguous classes increase as well resulting in less accuracies.

TABLE V
TOP-1, TOP-5 AND TOP-10 AVERAGE ACCURACY FOR POSE-BASED AND APPEARANCE-BASED MODELS ON THE WLASL-100, WLASL-300 AND WLASL-1000 DATASET

Type	Model	WLASL100			WLASL300			WLASL1000		
		top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10
Appearance based+ Backbone	I3D [15]	65.89	84.11	89.92	56.14	79.94	86.98	47.33	76.44	84.33
	TK-3D Convnet [46]	**77.55**	**91.42**	-	68.75	89..41	-	-	-	-
	Fusion3 [11]	75.67	86.00	90.16	68.30	83.19	86.22	56.68	79.85	84.71
Pose Based	Pose-GRU [15]	46.51	76.74	85.66	33.68	64.37	76.05	30.01	58.42	70.15
	Pose-TGCN [15]	55.43	78.68	87.60	38.32	67.51	79.64	34.86	61.73	71.91
	GCN-BERT [16]	60.15	83.98	88.67	42.18	71.71	80.93	-	-	-
	MOPGRU [47]	63.18	-	-	-	-	-	-	-	-
	SPOTER [17]	63.18	-	-	43.78	-	-	-	-	-
	SIGNGRAPH (ours)	72.09	88.76	**92.64**	**71.40**	**92.26**	**94.16**	**61.83**	**85.87**	**91.04**

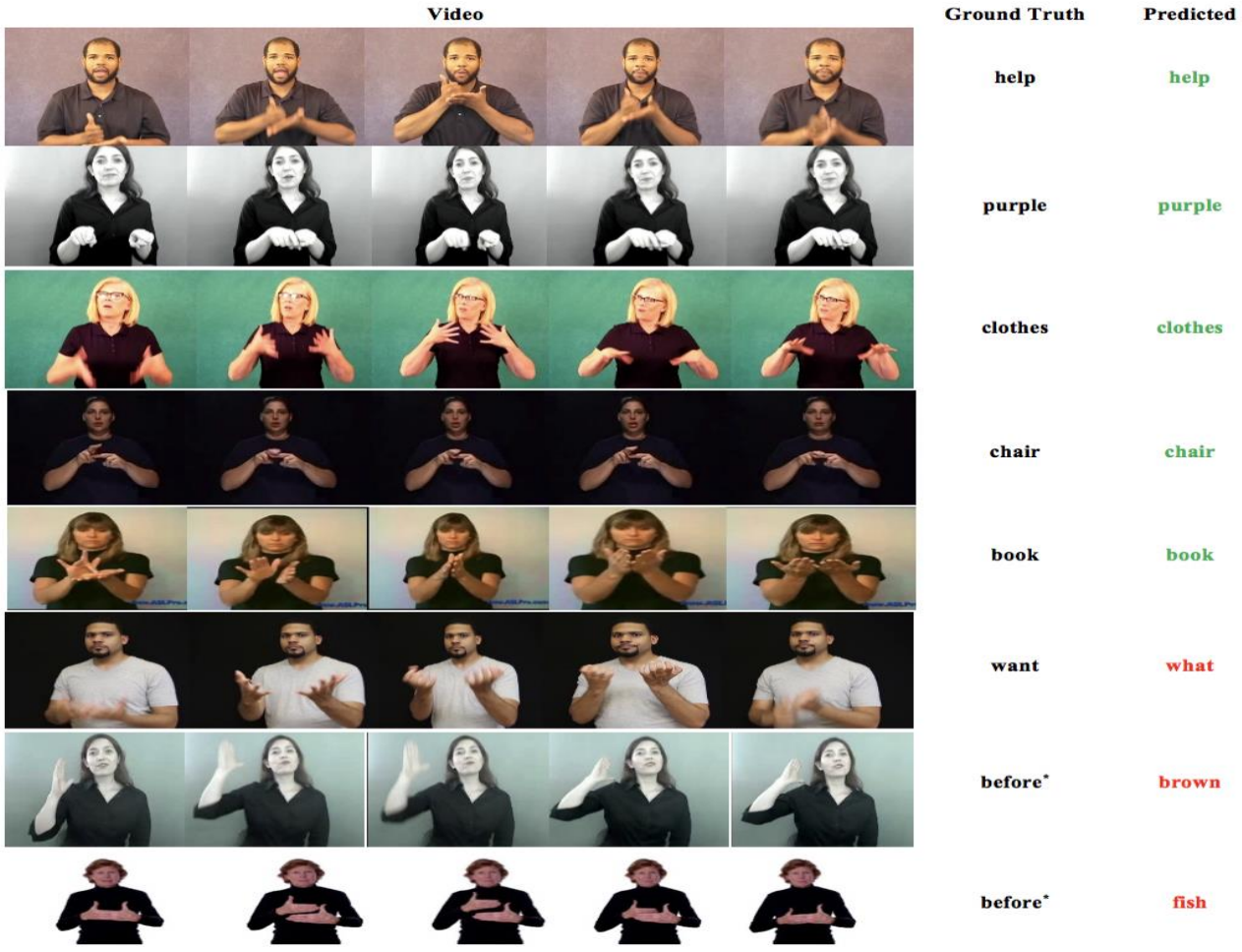

Figure 5. Examples of videos from WLASL along with their ground truth & predicted label

*last two rows represent the same word signed by two different signers in entirely different way

We have also tested our model on the LSA-64 dataset and results are reported in Table VI. We have compared our results on LSA-64 with the latest appearance-based, pose-based, and hybrid (pose + appearance) methods and our model outperformed all the methods by establishing SOTA test accuracy of 100%. Accuracies represented with * are obtained using cross validation over 5 repetitions.

TABLE VI
TOP-1 ACCURACY FOR POSE BASED AND APPEARANCE BASED MODELS ON THE LSA-64 DATASET

Type	Model	Top-1 Accuracy
Appearance based	LSTM+LDS [48]	98.09 ± 0.59 *
	DeepSign CNN [49]	96.00
	MEMP [50]	99.06
	I3D	98.91
Appearance + Pose	LSTM+DSC [51]	99.84± 0.19*
	ELM+MN CNN [33]	97.81
Pose Only	SPOTER [17]	100.00± 0 *
	MOPGRU [47]	99.92
	SIGNGRAPH (ours)	**100.00± 0***

D. EFFECT OF VOCABULARY SIZE

Vocabulary size in the dataset greatly impacts model performance. As presented in Table V, the models trained on datasets with smaller vocabulary sizes perform better as compared to larger ones.

To explain the impact of this factor, we have also tested the models trained on the dataset with larger vocabulary size to its smaller counterparts and results have been reported in Table VII. It can be seen in the table that for smaller vocabulary sizes the models are able to achieve higher accuracy. These experiments imply that classification becomes easier with the decrease in number of classes to be classified because of the decrease in the number of ambiguous classes.

V. ABLATION STUDY

Ablation study has been performed to analyze the influence of various components to system's performance. We have used the original ST-GCN model as a baseline. Bottleneck blocks and residual connections have been introduced to enhance the model performance and reduction in computational complexity and model size. Table VIII shows the results of the ablation study in terms of the baseline model, varying model depths, and residual connections. N represents the number of bottleneck blocks and R represents reduction rate. Thus, ResGCN-N39-R4 is a deeper architecture with a greater number of bottleneck blocks and a reduction rate of 4, whereas ResGCN-N21-R8 represents relatively a smaller number of bottleneck blocks and a reduction rate of 8. The results of the ablation study support our claim that using the residual connections and bottleneck structure improves the model's performance significantly.

VI. PERFORMANCE ANALYSIS

In this section, a comparative analysis of SIGNGRAPH (proposed architecture) and I3D (appearance-based) architecture is performed to assess the computational efficiency and performance of appearance and pose-based models. First, the number of model parameters are counted. For the WLASL-100 subset, SIGNGRAPH has only 0.62 million parameters, I3D has 12.4 million parameters, i.e., more than twenty times as much. Both models are assessed by their computational complexity in terms of their inference efficiencies and times. We have computed floating-point operations (FLOPs) during inference using Deepspeed [52] library's FLOP profiler. We have chosen a random batch of 32 videos from WLASL dataset and averaged the required FLOPs and inference time. Fig.6 shows all these performance attributes. The evaluations have been carried out on a single NVIDIA RTX-3080 GPU. On our system, SIGNGRAPH took 0.040 seconds to process each video and required 1.95M GFLOPs on average, Whereas the I3D took 0.50 seconds and required 5.22 GFLOPs on average.

To test our model's ability to learn more generalizable and robust representation, we have performed experiments by training our model on smaller training subsets and testing on a fixed set. We sampled the sizes of the training dataset and trained our model on these subsets. The learnt model is than tested on a fixed test set. We have conducted these experiments on LSA-64 because of its smaller size. We split the entire data into training and test set by an 80:20 ratio. Next, 5 SIGNGRAPH and I3D models have been trained, with a different split of training set each time.

TABLE VII
TOP-10 ACCURACY (%) FOR POSE BASED AND APPEARANCE BASED MODELS TRAINED(ROW) AND TESTED(COLUMN) ON DIFFERENT WLASL SUBSETS

	WLASL-100			WLASL-300			WLASL-1000		
Model	**I3D**	**TGCN**	**SIGNGRAPH (Ours)**	**I3D**	**TGCN**	**SIGNGRAPH (Ours)**	**I3D**	**TGCN**	**SIGNGRAPH (Ours)**
WLASL-100	89.92	87.60	**92.64**	-	-	-	-	-	-
WLASL-300	88.37	81.40	**92.56**	86.98	79.64	**94.61**	-	-	-
WLASL-1000	85.27	77.52	**90.70**	86.22	74.25	**92.51**	84.33	71.91	**91.04**

TABLE VIII
ABLATION STUDY ON WLASL-100, WLASL-300, AND WLASL-1000 DATASET SPLITS IN TERMS OF BOTTLENECK BLOCKS (MODEL DEPTH) AND RESIDUAL CONNECTIONS

Model	WLASL-100 (Accuracy %)	WLASL-300 (Accuracy %)	WLASL-1000 (Accuracy %)
ST-GCN (Baseline)	51.55	36.20	26.63
ResGCN-N21-R8	67.83	69.35	60.71
ResGCN-N21-R4	69.21	70.25	60.99
ResGCN-N39-R8	71.45	70.81	61.62
ResGCN-N39-R4 (SIGNGRAPH)	**72.09**	**71.40**	**61.83**

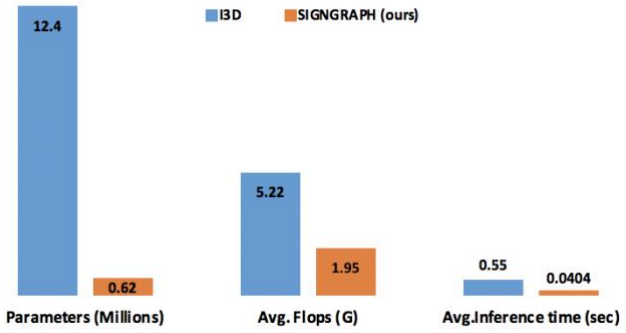

Figure 6. Relative comparison of the SIGNGRAPH and I3D model attributes: 1. Number of model parameters, 2. Average FLOPS 3. Average Inference time

We started training models with a 10% subset and went all the way up to the full data by adding 20% more data each time. To preserve the class distributions, training subsets were distributed uniformly. We evaluated the trained model on the original fixed test subset. The results are shown in Fig.7. SIGNGRAPH achieved an accuracy of 74.99% even when trained only on 10% of the training set, whereas I3D model lagged behind with 45.47% accuracy. SIGNGRAPH was able to achieve an accuracy of 100% when trained only on 70% subset whereas I3D model achieved maximum accuracy of 98.91% when full training set was used.

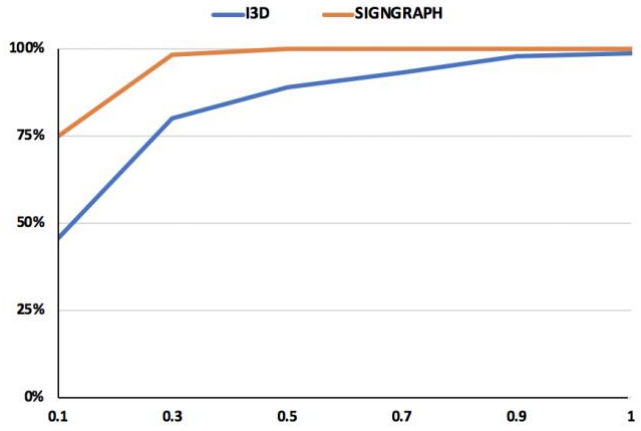

Figure 7. Top-1 accuracies of SIGNGRAPH and I3D models when trained on five gradually increasing subsets of training set and tested on a fixed 20% split for LSA-64.

The reason for this behavior can be explained as; I3D model requires learning general concepts like human body mechanics for semantic decoding of sign language. When the model is trained on smaller data sizes, learning becomes even harder, thus degrading the model's performance. Hand crafted body and hand pose information fed to SIGNGRAPH already contains enough information for such decoding resulting in higher accuracies even when smaller number of data samples are used. Considering model size, data size, computational requirements for inference, and speed, these experiments clearly demonstrate the superiority of SIGNGRAPH for SLR as opposed to appearance-based approaches like I3D.

VII. CONCLUSION

This paper proposes a pose-based, residual graph convolution network, to the task of isolated SLR. Our proposed architecture, SIGNGRAPH, uses 2D hands and body skeleton pose as input features. By considering the inherent graph structure of the pose, sign information is extracted. Previously proposed architectures to tackle the problem are less accurate and computationally heavy. Experiments conducted on the three different subsets of the latest American sign language database WLASL and Argentinian sign language LSA-64 show SOTA results in pose-based and appearance-based SLR methods while reducing computational complexity. The ablation study also shows the importance of residual connections and bottleneck structures in improving the model's performance. A performance comparison of the proposed architecture with the appearance-based methods proves that our proposed architecture is significantly less demanding and generalizes well. In the future, we plan to extend the proposed architecture by using multi branches of hand-crafted features and by introducing an attention mechanism to learn the most significant motion patterns for efficient SLR.

ACKNOWLEDGMENT

The authors would like to acknowledge the role, contribution, guidance and support provided by DeafTawk (https://www.deaftawk.com/) in terms of domain knowledge for successful completion of this research work.

REFERENCES

[1] M. J. http://wfdeaf.org/our-work/. (accessed 30 January 2020, 2020).
[2] C. Valli, Lucas, C., *Linguistics of American Sign Language: An Introduction. Gallaudet* University Press, 2000.
[3] D. Brentari, *Sign Language Phonology*. Cambridge University Press 2019.

[4] J. S. Supančič, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: methods, data, and challenges," *International Journal of Computer Vision*, vol. 126, no. 11, pp. 1180-1198, 2018.
[5] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural computing and applications*, vol. 32, no. 12, pp. 7957-7968, 2020.
[6] A. Tagliasacchi, M. Schröder, A. Tkach, S. Bouaziz, M. Botsch, and M. Pauly, "Robust articulated-icp for real-time hand tracking," in *Computer graphics forum*, 2015, vol. 34, no. 5: Wiley Online Library, pp. 101-114.
[7] T. G. Zimmerman, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvill, "A hand gesture interface device," *ACM Sigchi Bulletin*, vol. 18, no. 4, pp. 189-192, 1986.
[8] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.
[9] P. Das and A. Ortega, "Symmetric sub-graph spatio-temporal graph convolution and its application in complex activity recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 3215-3219.
[10] Z. Huang, X. Shen, X. Tian, H. Li, J. Huang, and X.-S. Hua, "Spatio-temporal inception graph convolutional networks for skeleton-based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2122-2130.
[11] A. A. Hosain, P. S. Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka, "Hand pose guided 3d pooling for word-level sign language recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3429-3439.
[12] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10863-10872.
[13] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3383-3393.
[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291-7299.
[15] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459-1469.
[16] A. Tunga, S. V. Nuthalapati, and J. Wachs, "Pose-based sign language recognition using gcn and bert," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 31-40.
[17] M. Boháček and M. Hrúz, "Sign Pose-based Transformer for Word-level Sign Language Recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 182-191.
[18] P. C. Badhe and V. Kulkarni, "Indian sign language translator using gesture recognition algorithm," in *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, 2015: IEEE, pp. 195-200.
[19] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: IEEE, pp. 2961-2968.
[20] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research*, vol. 13, pp. 2205-2231, 2012.
[21] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293-1301.
[22] G. D. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in *European Conference on Computer Vision*, 2014: Springer, pp. 595-607.
[23] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *2016 IEEE international conference on multimedia and expo (ICME)*, 2016: IEEE, pp. 1-6.
[24] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 430-439, 2018.
[25] O. M. Sincan, A. O. Tur, and H. Y. Keles, "Isolated sign language recognition with multi-scale features using LSTM," in *2019 27th signal processing and communications applications conference (SIU)*, 2019: IEEE, pp. 1-4.
[26] H. Luqman and E.-S. M. El-Alfy, "Towards hybrid multimodal manual and non-manual Arabic sign language recognition: MArSL database and pilot study," *Electronics*, vol. 10, no. 14, p. 1739, 2021.
[27] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *arXiv preprint arXiv:1812.01053*, 2018.
[28] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299-6308.
[29] Z. Liu, L. Pang, and X. Qi, "MEN: Mutual Enhancement Networks for Sign Language Recognition and Education," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[30] Y. C. Bilge, R. G. Cinbis, and N. Ikizler-Cinbis, "Towards zero-shot sign language recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[31] H. Park, Y. Lee, and J. Ko, "Enabling real-time sign language translation on mobile platforms with on-board depth cameras," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1-30, 2021.
[32] C. C. dos Santos, J. L. A. Samatelo, and R. F. Vassallo, "Dynamic gesture recognition by using CNNs and star RGB: A temporal information condensation," *Neurocomputing*, vol. 400, pp. 238-254, 2020.
[33] J. Imran and B. Raman, "Deep motion templates and extreme learning machine for sign language recognition," *The Visual Computer*, vol. 36, no. 6, pp. 1233-1246, 2020.
[34] O. M. Sincan and H. Y. Keles, "Using Motion History Images with 3D Convolutional Networks in Isolated Sign Language Recognition," *IEEE Access*, vol. 10, pp. 18608-18618, 2022.
[35] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413-3423.
[36] O. M. Sincan, J. Junior, C. Jacques, S. Escalera, and H. Y. Keles, "Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3472-3481.
[37] F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. Rosete, "LSA64: an Argentinian sign language dataset," in *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*. 2016.
[38] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: IEEE, pp. 2200-2207.
[39] O. M. Sincan and H. Y. Keles, "Autsl: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340-181355, 2020.
[40] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693-5703.
[41] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349-3364, 2020.
[42] Google. "MediaPipe Holistic." https://google.github.io/mediapipe/solutions/holistic. (accessed 13th December 2022, 2022).

[43] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1625-1633.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[46] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring cross-domain knowledge for video sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6205-6214.
[47] B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K.-H. Park, and J. Kim, "An integrated mediapipe-optimized GRU model for Indian sign language recognition," *Scientific Reports*, vol. 12, no. 1, pp. 1-16, 2022.
[48] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Sign language recognition based on hand and body skeletal data," in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2018: IEEE, pp. 1-4.
[49] J. A. Shah, "Deepsign: A deep-learning architecture for sign language," 2018.
[50] X. Zhang and X. Li, "Dynamic gesture recognition based on MEMP network," *Future Internet*, vol. 11, no. 4, p. 91, 2019.
[51] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "A deep learning approach for analyzing video and skeletal features in sign language recognition," in *2018 IEEE international conference on imaging systems and techniques (IST)*, 2018: IEEE, pp. 1-6.
[52] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3505-3506.

Neelma Naz received the M.S. degree in Electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2014, where she is currently pursuing the Ph.D. degree in Robotics and Intelligent Machine Engineering with the from School of Mechanical and Manufacturing Engineering.
Her current research interests include computer vision, pattern recognition, machine learning and control systems.

Hasan Sajid received the B.S. degree in mechatronics engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from University of Kentucky, USA, in 2014 and 2016, respectively. He is currently an Associate Professor at Department of Robotics & AI, NUST and Scientific Director at National Center for Artificial Intelligence. He has expertise in the areas of computer vision, machine learning and deep learning. His research interests include speech and text recognition, video analytics and application of AI in healthcare, crowd and traffic domains. He has 25+ high impact peer reviewed publications and won fundings of more than 100 M. He was a recipient of the U.S. State Department Fulbright Scholarship.

SARA ALI received her PhD degree in Robotics and Intelligent Machine Engineering from School of Mechanical and Manufacturing Engineering, National University of Sciences and Technology (NUST), Pakistan in 2020. She did her master's in research from Middlesex University, London, UK in 2013 on faculty development program (FDP). Currently she is an Assistant Professor in Robotics and Intelligent Machine Engineering Department, NUST, Pakistan. Her research interest includes Human-Robot Interaction, Sensor Systems, Interactive Robotics, Virtual Reality, and Human-Machine Interaction. She is the Principal Investigator (PI) of 2 main labs i.e., Intelligent Field Robotics Lab (IFRL) and Human-Robot Interaction (HRI) lab at National University of Sciences and Technology. She has published over 40 research articles in international peer-reviewed journals and conferences including citation from Nature. She has also been appointed as session chair and member of Scientific Advisory Board (SAB) in several prestigious international conferences. She has also authored a book titled "Introducing Therapeutic Robotics for Autism" in UK published by Emerald Publishing Ltd. She has also been awarded a national patent on Brain Imaging Tool for Medical Diagnosis.

Osman Hasan received his BEng (Hons) degree from the University of Engineering and Technology, Peshawar Pakistan in 1997, and the MEng and PhD degrees from Concordia University, Montreal, Quebec, Canada in 2001 and 2008, respectively. Before his PhD, he worked as an ASIC Design Engineer from 2001 to 2004 at LSI Logic. He worked as a postdoctoral fellow at the Hardware Verification Group (HVG) of Concordia University for one year until August 2009. Currently, he is Pro-Rector (Academics) at National University of Science and Technology (NUST), Islamabad, Pakistan. He is the founder and director of System Analysis and Verification (SAVe) Lab at NUST, which mainly focuses on the design and formal verification of energy, embedded and e-health related systems. He has received several awards and distinctions, including the Pakistan's Higher Education Commission's Best University Teacher (2010) and Best Young Researcher Award (2011) and the President's gold medal for the best teacher of the University from NUST in 2015. Dr. Hasan is a senior member of IEEE, member of the ACM, Association for Automated Reasoning (AAR) and the Pakistan Engineering Council.

Muhammad Khurram Ehsan received his Ph.D. degree in engineering with specialization in statistical signal processing and the M.S. degree in electrical communication engineering from the University of Kassel, Germany, in 2016 and 2010, respectively. He has been an Associate Professor, Faculty of Engineering, Bahria University, Pakistan, since July 2022. He has been a Visiting Lecturer, Faculty of Electrical Engineering and Computer Science, University of Kassel, Germany, since July 2017. His research interests include statistical modeling, data analysis, and cognitive radio systems.