IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 October 2023, accepted 11 December 2023, date of publication 15 December 2023,
date of current version 22 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3343404

RESEARCH ARTICLE

Dynamic Korean Sign Language Recognition Using Pose Estimation Based and Attention-Based Neural Network

JUNGPIL SHIN[1], (Senior Member, IEEE), ABU SALEH MUSA MIAH[1], KOTA SUZUKI[1], KOKI HIROOKA[1], AND MD. AL MEHEDI HASAN[2], (Member, IEEE)

[1]School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-0006, Japan
[2]Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh

Corresponding author: Jungpil Shin (jpshin@u-aizu.ac.jp)

This work was supported by the Competitive Research Fund of The University of Aizu, Japan.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Sign language recognition is crucial for improving communication accessibility for the hearing impaired community and reducing dependence on human interpreters. Notably, while significant research efforts have been devoted to many prevalent languages, Korean Sign Language (KSL) remains relatively underexplored, particularly concerning dynamic signs and generalizability. The scarcity of KSL datasets has exacerbated this limitation, hindering progress. Furthermore, most KSL research predominantly relies on static image-based datasets for recognition, leading to diminished accuracy and the inability to detect dynamic sign words. Furthermore, most KSL research predominantly relies on static image-based datasets for recognition, leading to diminished accuracy and the inability to detect dynamic sign words. Additionally, existing KSL recognition systems grapple with suboptimal performance accuracy and heightened computational complexity, further emphasizing the existing research gap. To address these formidable challenges, we propose a robust dynamic KSL recognition system that combines a skeleton-based Graph Convolution network with an attention-based neural network, effectively bridging the gap. Our solution employs a two-stream deep learning network to navigate the intricacies of dynamic signs, enhancing accuracy by effectively handling non-connected joint skeleton features. In this system, the first stream meticulously processes 47 pose landmarks using the Graph Convolutional Network (GCN) to extract graph-based features. These features are meticulously refined through a channel attention module and a general CNN, enhancing their temporal context. Concurrently, the second stream focuses on joint motion-based features, employing a similar approach. Subsequently, these distinct features from both streams are harmoniously integrated and channelled through a classification module to achieve precise sign-word recognition. A significant contribution of our work lies in creating a novel KSL video dataset, addressing the scarcity of data in this domain. This dataset comprises comprehensive information, including skeletal data from 47 joint skeleton points and details from both hands, body, and facial expressions. Our dataset aims to fill a critical gap in KSL research and provides a solid foundation for more extensive and inclusive studies in the field. Through this innovative approach, we aim to contribute significantly to the field of KSL recognition, filling the gaps in dynamic sign recognition and bolstering the accessibility of sign language communication within the Korean hearing impaired community and beyond. Our evaluations on a benchmark KSL-77 dataset and our proprietary lab dataset resulted in recognition accuracies of 99.87% and 100%, respectively. These results highlight the superiority of our model in the KSL recognition domain, outperforming existing models in terms of accuracy and computational efficiency.

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

INDEX TERMS Dynamic hand gesture recognition, Korean sign language (KSL), graph convolutional network (GCN), general convolutional neural network (GCNN), machine learning, hand skeleton points, deep learning.

ABBREVIATIONS

KSL	Korean Sign Language.
CNN	Convolutional Neural Network.
GCN	Graph Convolutional Network.
ASL	American Sign Language.
PCA	Principal component analysis.
HMM	Hidden Markov model.
ANN	Artificial neural networks.
FCA	Fuzzy classification algorithms.
KNN	K-nearest-neighbor.
DGSTA	Dynamic Graph-based Spatial-Temporal Attention.
GSTCAN	Graph-based Spatial-Temporal Convolution and Attention Network.
ASGCN	Spatial Graph Convolutional Network.
GSCAN	Graph Spatial Convolution and Attention Network.

I. INTRODUCTION

Sign Language recognition (SLR) is a crucial communication medium for the hearing impaired community, providing a way to interact without relying on traditional spoken languages. Worldwide, the World Health Organization (WHO) reports that approximately 430 million individuals are affected by hearing loss [1], [2], [3], [4]. To facilitate communication between the hearing impaired and non-hearing impaired populations using sign language, it is essential for both communities to learn sign language. However, learning sign language is challenging due to the specific motions required to convey various signs. It necessitates individuals to acquire a significant number of distinct gestures separate from spoken language. Additionally, sign language varies significantly from one country to another and even within the same country, resulting in entirely different sets of signs. For instance, Korean Sign Language (KSL) is entirely distinct from American Sign Language (ASL) and Bangla Sign Language (BSL) [5]. Furthermore, sign language can vary within regions, sometimes leading to the existence of multiple sign languages within a single spoken language. For example, British Sign Language features different gestures from ASL, despite both being used alongside the English language. In such situations, only human translators can facilitate communication between the hearing impaired and non-hearing impaired communities for Korean people. However, obtaining a human translator for Korean individuals can be challenging due to cost and efficiency concerns [6]. In this context, an automatic sign language translator becomes the only viable solution for communication between the hearing impaired and non-hearing impaired communities. A considerable amount of research has been conducted on sign language recognition for various languages, including English, Arabic, Turkish, Indian, and others [7], [8], [9], [10]. However, Korean sign language(KSL) recognition has seen limited development due to the absence of a comprehensive KSL dataset. Kim et al. introduced a dynamic KSL recognition system using a fuzzy neural network, recording 31 KSL alphabet signs with a hand glove system [11]. While effective, this system relied on hardware and sensors, leading to issues related to portability and high costs. In response, researchers have shifted their focus towards vision-based systems, utilizing webcams or cameras for increased portability and cost-effectiveness. Other researchers proposed a vision-based Korean sign language classification model employing an ensemble artificial neural network (ANN) [6]. They utilized ten labels and 1,500 sample images captured with a high-quality camera, achieving an accuracy of 97.4%, primarily focusing on finger spelling signs. However, the limitation of their approach is the relatively small number of signs considered, which may not suffice for a real-life KSL recognition system. To address this limitation, Yang et al. collected a large-scale Korean sign word dataset and achieved 79.80% accuracy using a deep learning network [12]. To further enhance performance accuracy, Shin et al. introduced a KSL recognition system employing a multi-branch transformer and a general CNN-based model [5]. However, their use of pixel-based images as input differs from existing KSL recognition systems that rely on RGB image-based approaches, potentially leading to performance challenges related to various backgrounds, partial occlusion, computational complexity, and varying lighting conditions. To address the complexity of pixel-based issues, Ko et al. released a KSL dataset and presented a KSL translation model designed to extract 2D human pose key points [13]. Nevertheless, they encountered challenges related to lower performance accuracy and high computational complexity.

Additionally, most existing research focuses on recognizing sign language using still images and does not consider the detection of dynamic sign words. Notably, there is a lack of research in the development of a dynamic KSL recognition system [14].

Moving forward, while advanced research combining computer vision, robotics, and natural language processing has made significant strides in the evolution of sign language recognition (SLR), there remains a conspicuous gap in dynamic KSL recognition. Existing studies tend to rely heavily on hardware and sensor-based systems [12], while those oriented towards static KSL recognition have limitations [6]. Furthermore, the field of skeleton-based video classification research in the context of KSL remains largely unexplored, highlighting a substantial research gap [13]. Recently, some researchers have employed Graph Convolutional Neural Networks (GCNs) for skeleton-based dynamic action recognition, as seen in works such as [15], GSTCAN [16], GSTCAN [17], ASGCN [14], and GSCAN [18].

Among these, the study by Shi et al. utilized two-stream GCN networks based on joint and joint motion skeleton-based information in an attempt to improve performance accuracy [14]. However, these models fell short of achieving high accuracy for KSL recognition. The primary drawback lies in their failure to consider non-connected skeleton joints and joint motion features, which could potentially enhance performance accuracy in KSL recognition. In light of these challenges and gaps in existing research, our proposed solution involves joint skeleton-based dynamic KSL recognition using two-stream deep learning networks. Each stream incorporates the Graph Convolutional Network (GCN) and an attention-based general neural network approach. To overcome these challenges, we proposed a joint skeleton-based KSL recognition using two two-stream deep learning networks where each stream is constructed with the Graph convolutional network(GCN) and attention-based general neural network approach. Our principal contributions encompass:

Our principal contributions encompass:

- **Novelty:** We've created a unique skeleton-based KSL video dataset featuring 47 joint skeleton points, including details from both hands, the body, and facial expressions. This dataset was carefully curated, considering diverse backgrounds and environmental conditions during data collection. By utilizing MediaPipe estimation to extract skeletal data from videos, we've successfully addressed challenges like differing backgrounds, partial obstruction, computational requirements, and varying lighting conditions. This dataset not only bridges a critical research gap but also establishes a robust basis for more extensive and inclusive studies in the realm of KSL recognition.
- **Methodological Innovation:** We've introduced a system designed to excel in recognizing dynamic KSL words through video classification, effectively surpassing the limitations of conventional static image-based methods. Within our system, we've crafted a dual-stream neural network that integrates the Graph Convolutional Network (GCN) with channel attention and a general deep learning model. Our primary approach entails constructing a graph and feeding it into the GCN module to generate graph-based features. To address issues related to non-connected joint skeleton data, we've enhanced these features using channel attention. Subsequently, we've further refined the features through a standard CNN module to enrich their temporal context. Finally, we've concatenated the features from both streams and input them into a classification module for sign language recognition.
- **Empirical Validation:** Through extensive experimentation conducted on both our newly established skeleton-based KSL dataset and a standard benchmark KSL-77 dataset, we've achieved remarkable accuracy rates of 100.00% and 99.87%, respectively. These results unequivocally demonstrate the effectiveness of our approach within the field of KSL recognition. The data and code have been uploaded to the following link:https://github.com/musaru/KSL.

The structure of the paper we organized is as follows: Related work described in Section II. Section III demonstrates the proposed methodology and explains the dataset, feature values, and classification method. Section IV describes the results, including the optimal parameter values and a comparison between systems with and without variation features shown V . Finally, Section VI describes the discussion, and the conclusion section follows this.

II. RELATED WORK

Static sign language recognition, continuous sign language recognition and dynamic sign language recognition are different categories of sign language. For decades, several kinds of machine learning and deep learning algorithms have been proposed for EMG, ECG, Image-based emotion, activity, and sign language recognition systems [5], [7], [8], [10], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. Every country has a different sign language, so developing SLR systems for the specific sign language has been attempted worldwide. In the Bengali sign language case, researchers developed the sign language recognition system with machine learning and deep learning algorithm [3], [4], [29]. Pitsikalis et al. proposed an SLR method using the Hidden Markov model (HMM), and they collected 961 images with a Kinect TM depth camera, which has depth channels and RGB channels [30]. However, their model achieved fine performance, while the HMMs model has a non-discriminatory learning feature, which occasionally misses the input from the alternative class. Moreover, their architecture cannot be placed into commercial use. Ong et al. employed the sequential pattern tree boosting algorithm for their multi-class sign language classification model, where they aimed to extract hand motion features from the sequential images [31]. Their model performed better than the HMM model, and the validation achieved 93.00% and 88.00% accuracy for the Greek and German sign language, respectively. Almeida et al.applied a phonological model to decompose the RGB-D image for feature extraction, and then they used a support vector machine (SVM) for the classification and reported 80.00% accuracy for their own dataset [32]. Fatimi et al. employed an artificial neural network (ANN) and SVM for an ASL recognition model, which had a higher accuracy score than HMM and SVM models [33]. Lee et al. employed the SVM for several wearable hand devices when their model performed 98.2% accuracy. In other cases, 98 % accuracy was yielded with the Chinese sign language (CSL) recognition model applied to the fuzzy network. Moreover, an ASL recognition model for ten terms of 0-9 used KNN, LDA and SVM, reaching 93.79 % score at best. Na et al. proposed a KSL recognition system with triaxial accelerometer signals, and after using SVM, they achieved 92.00% accuracy [34]. Kim et al. proposed a dynamic KSL recognition system using a fuzzy neural

network system where they recorded the 31 KSL alphabet using hand gloves system [11] In the same way, various conventional machine learning algorithm-based SLR systems proposed by many researchers, such as hidden conditional random field (HCRF), HMM, and random decision forest (RDF) which produced good performance scores with a small datasets [35], [36]. However, some problems remained, such as the heavy computational complexity of a large dataset, which makes the classification performance slower. As a solution for these problems, some researchers proposed an ANN algorithm for sign language recognition [37], [38]. Kim et al. proposed a Korean sign language classification model employing ensemble ANN [6]. They prepared the ten labels and 1500 samples and focused on finger spelling signs; consequently, it performed with 97.4% accuracy. Ko et al. released a KSL dataset and presented a KSL translation model for extracting 2D human pose key points [13]. For some years, the advances in artificial intelligence and computing technologies switched to the deep learning-based model. Al-Hammadi et al. designed a 3-dimensional CNN by including single and parallel branch-based methods to develop an SLR system where their method achieved good performance for three sign language datasets [39]. In their evaluations, they yielded accuracy scores of 84.38%, 34.9%, and 70% for datasets consisting of 40, 23, and 10 classes under the singer-independent condition. Moreover, they improved by +10% under the signer-dependent condition, where their work was better than similar previous cases. Sincan et al. developed a hybrid method by combining a long short-term memory(LSTM) with the CNN, aiming to generate an attention-based feature and feature pooling model(FPM) [15]. Their model features an attention module to expedite solving the convergence point, where a high accuracy score was performed in the evaluation for Turkish sign language. Yuan et al. developed an SLR system using LSTM and Deep CNN, and then they evaluated their model with ASL and CSL dataset [15]. While their model overcame the gradient vanishing and overfitting problem, they need to solve the long-distance dependency problems, which are not resolved in their method. Aly et al. developed an Arabic sign language recognition system using a deep bi-direction LSTM (BiLSTM) classification method through a self-organization map-based features [40]. Some researchers applied deep learning models such as 2DCNN, 3DCNN and LSTM architecture for the skeleton-based SLR system [41]. For more than a few years, such as CNN with attention, VGGNet and AlexNet, several existing CNN architectures have been employed to overcome particular issues [42], [43], [44]. These architectures mainly consisted of deep learning layers such as convolution, dropout, and pooling. The multiple paths of that layer prove their effectiveness in GoogleNet and InceptionNet [45], [46]. ResNet enhanced generalization performances by incorporating shortcut connections every two layers to the base network [47]. The mentioned method is not effective for the non-connected skeleton point for skeleton-based gesture recognition. An attention-based architecture module was applied to an operator between adaptable modalities as a solution to the existing issues [48]. Transformers have frequently been employed for novel vision tasks since the remarkable advance of natural language processing with it [42], [47], [49]. ViT is a well-known transformer architecture directly adapted from other research domains to the computer vision research domain for the sign language classification task. However, it is required a huge amount of dataset to perform high accuracy score [50], when a researcher presented DeiT, which overcomes the huge dataset issue and enables to process train task with fine efficiency [51]. After appearing ViT, an enhanced transformer T2T-ViT was introduced, which converts the neighbouring tokens into individual tokens recursively [52]. The problem with ViT is that it sometimes loses some potential information because it relies on only the patch sequence. Then, a TNT transformer, which has inner and outer blocks by incorporating both pixel-level information and patch-level details, was proposed. PVT, CvT and CPVT can be integrated into the general deep learning layer and transformer to solve the long-range dependencies issue, whereas the combined utilization did not perform well [53], [54], [55]. The CMT technique was proposed to overcome the combination inefficiency issue, consisting of a transformer, four stages, and the CNN. However, the disadvantage is that the CMT of mixed short-term and long-term dependency in each stage causes rising computational complexity. Additionally, while numerous deep learning models based on transformers have been developed, adequate performance has not yet been yielded [56]. Recently, shin et al. used a multi-branch CMT-based transformer and a general CNN-based model to develop the KSL recognition system. The main drawback of the work is that they used pixel-based images as input for the dataset [5]. They still face problems such as lower performance accuracy and high computational complexity. In addition, we did not find a skeleton-based dynamic KSL recognition system yet. Recently, skeleton-based GCN and attention modules have proven excellent in other sign language recognition work [1], [14], [18], [56]. However, most of the KSL research work is related to static KSL recognition where no motion is included, and the dynamic nature of the sign is not considered. A limited number of dynamic KSL recognition research works have been done, mostly using hardware and sensor-based systems, which still have many drawbacks. Thus, it is urgent to develop a vision-based dynamic KSL recognition model to recognize KSL from video or dynamic data. In addition, KSL recognition is still a bit challenging work because of the diversity of the signs coming from human gestures. In addition, arbitrary-view and dynamic signs come from the multiple-camera viewpoint. In addition, no skeleton-based KSL recognition work has been done yet. To overcome the problems, we developed a skeleton-based Graph convolution and attention-based general neural network to achieve

satisfactory accuracy and efficiency in the KSL recognition systems.

III. DATASET DESCRIPTION

Also, KSL is one of the most used people using this language, but little research has been done, and few datasets are available online. We have only one dataset online, the large-scale Korean sign language (KSL) dataset [12]. In the study, we worked with the dataset to evaluate the proposed model: large-scale KSL dataset described in Section III-A. We create a new KSL selection data set, which we described in Section III-B

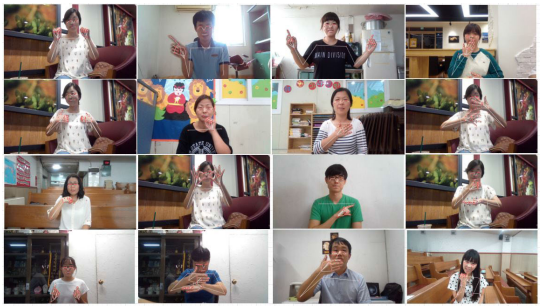


FIGURE 1. Example of KSL image from 77 class KSL dataset.

A. LARGE KSL- DATASET

Large KSL is one of the most usable benchmark datasets for Korean sign language recognition work. This is the first large-scale KSL dataset for the Korean sign language, one of the most used languages in the world. This dataset was recorded from 20 people for 77 sign words. Although many daily activity words are available in the Korean sign language, they tried to include the most used sign word in their data. During the data collection, they used 17 diverse backgrounds and locations for multiple signers where they considered facial expressions besides hand gestures. They considered the various distances and angles for collecting the actual scenario to account for the real-life situation. There are 1229 videos in their dataset laterally; it generated 112564 frames from the videos. They considered a 30 ps frame rate for converting the video dataset into frames, and they discarded a few initial frames and a few end frames to keep the actual information by removing the noise and empty frames. For hand pose extraction, we used the media pipe approach, which took the RGB video as an input and produced the joint skeleton point for the hand and body pose key points, which fed the network as an input dataset. Increasing the privacy of the signer and reducing the computational complexity are the main purposes of using joint skeleton information here instead of pixel-based images. The main goal of the skeleton dataset is to increase the privacy and efficiency of the system by avoiding the exact pixel information and scenarios. Figure 1 visualized the skeleton points of this dataset.

FIGURE 2. Proposed korean sign language word.

B. PROPOSED LAB DATASET

To overcome the unavailability of the KSL dataset, we created a word and signer-independent KSL dataset. To make this dataset, we selected the 20 most significant words Korean people use for daily activities. The name of these words dataset included thanks, love, okay, no, happy, sorry, hello, shame, late, regrettable, meet, yes, help me, effort, give, welcome, what, by, why and who [6], [12], [55]. Although the existing KSL dataset included 77 KSL words to make their dataset, that number is also good for expressing ideas, thoughts, expressions and requirements to other people. However, not all 77 words are frequently used for their daily activities. We investigated that some of them are mostly used to do daily activities, but some other KSL words can be used more frequently to express a human's basic needs and thoughts compared to the included 77 words. In the proposed new dataset, we also included the five most usable sign words from the existing large-scale KSL dataset [55]. We considered more than 15 KSL words with high significance to use in daily life for the deep and mute Korean people. The five sign words that we selected from the previous dataset are as follows: no, thanks, what, sorry, and who. We selected more than 15 words by studying and seeing the problem with the current sign languages. The study mainly considered the skeleton point instead of the image pixels. The sample image of the dataset is visualized in Figure 2. As an environment, we used a webcam-based RGB camera to record the 20 words for the KSL dataset. The sample RGB picture with skeleton mapping for the same picture is shown in Figure 4. To record the dataset, we used 4-second videos with 120 frames, most of which were people, and 30 people willingly connected with us to record the dataset. Moreover, The background of the proposed KSL video dataset recorded the various scenarios as possible with a natural background.

IV. PROPOSED METHODOLOGY

Figure 3 demonstrates the proposed workflow architecture. Research on KSL recognition primarily focuses on using still images for sign language recognition, often failing to detect dynamic sign words. In addition, the existing system may fail to achieve high-performance accuracy with the benchmark dataset because of the unavailability of the

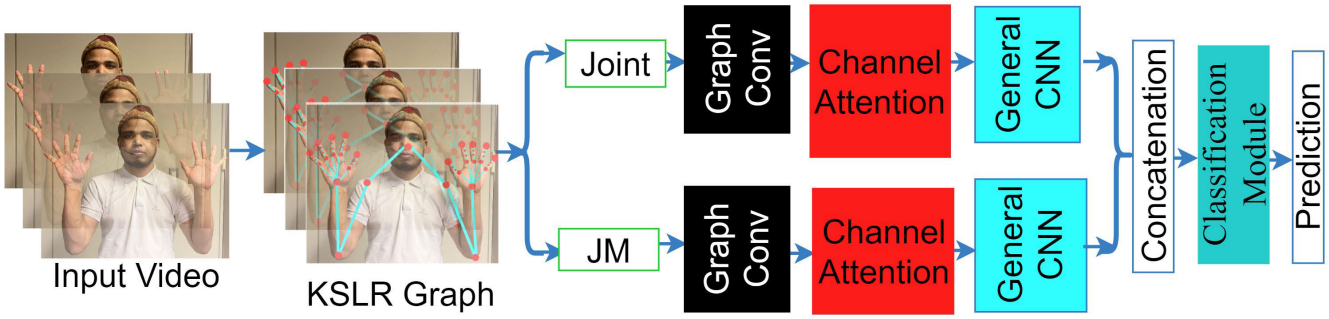


FIGURE 3. Proposed korean sign language word recognition architecture.

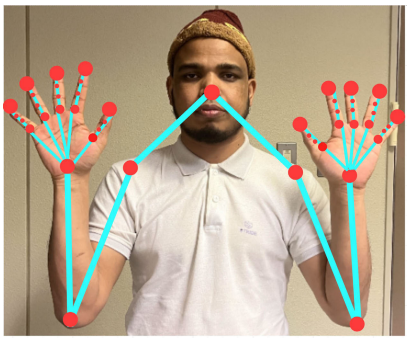


FIGURE 4. Skeleton key points for the proposed model.

training dynamic KSL dataset. Recently, some researchers have employed Graph Convolutional Neural Networks (GCNs) for skeleton-based dynamic action recognition, as seen in works such as DGSTA, [16], GSTCAN [17], ASGCN [14], and GSCAN [18]. Among these, the study by Shi et al. utilized two-stream GCN networks based on joint and joint motion skeleton-based information in an attempt to improve performance accuracy. However, these models fell short of achieving high accuracy for KSL recognition. The primary drawback lies in their failure to consider non-connected skeleton joints and joint motion features, which could potentially enhance performance accuracy in KSL recognition. In light of these challenges and gaps in existing research, our proposed solution involves joint skeleton-based dynamic KSL recognition using two-stream deep learning networks. Each stream incorporates the Graph Convolutional Network (GCN) and an attention-based general neural network approach. To overcome these challenges, we proposed a joint skeleton-based dynamic KSL recognition using two two-stream deep learning networks where each stream is constructed with the Graph convolutional network(GCN) and attention-based general neural network approach. To do this, we derived a new skeleton dataset from the recorded videos using the Mediapipe pose estimation method, overcoming the mentioned challenges. This estimation captures skeletons from both hands, as well as facial and body points, to consider the emotional information besides the only hand gesture information. We then analyzed motion in the skeleton-based dataset. Our proposed two-stream neural network architecture employs a Graph Convolutional Network (GCN) and an attention-based neural approach. The first stream uses Mediapipe to extract 47 pose landmarks, which GCN processes to create graph-based features. These are further refined by a channel attention module and a general CNN to enhance temporal context. The second stream captures joint motion data using a similar feature extraction process. Features from both streams are combined and fed into a classification module for sign language recognition. Detailed descriptions of our methodology are provided in subsequent sections.

A. POSE ESTIMATION

In the study, we considered a hand skeleton point instead of a pixel-based image. One of the main reasons is to hide the human hand's information to protect privacy and security. The idea is the skeleton information does not contain the details of visual information such as texture or skin colour. This idea allows a person to anonymise their identity, and we will not record or store the hand's appearance. The texture of the hands is relevant to preserving crucial biometric systems and so on. Moreover, fingerprints and palm prints, which are very sensitive information for a person, are not required to be exposed to the skeleton dataset. Skeleton dataset can protect this personal information from unauthorized access to their biometric data. Another advantage of the skeleton dataset is that it needs significantly less storage of the hand gesture data. Shortly, using the skeleton dataset, we offered a secured hand gesture recognition system that hides personal information while enabling secure privacy and authentication protection. The study used a media pipe system to extract the skeleton joint from the video dataset. We collected 21 skeletons from the left hand, 21 skeleton points from the right and points from the body. In total, there are 47 skeleton key points we extracted here [18], [57].

B. CAPTURING JOINT MOTION

In the study, we included static and dynamic signs for the KSL sign word, although existing research developed dynamic signs in the static system. Motion is a very effective feature for dynamic gestures in terms of alignment and movement

TABLE 1. The list of the key points used in the study.

Sr No	Hand Pose No	Left Hand Pose Name	Sr No	Body Pose No	Body Pose Name
1	0	Wrist	25	12	Right shoulder
2	1	Thumb CMC	26	14	Left elbow
3	2	Thumb MCP		Hand Pose No	Right Hand Pose Name
4	3	Thumb IP	27	0	Wrist
5	4	Thumb TIP	28	1	Thumb CMC
6	5	Index CMC	29	2	Thumb MCP
7	6	Index MCP	30	3	Thumb IP
8	7	Index IP	31	4	Thumb TIP
9	8	Index TIP	32	5	Index CMC
10	9	Middle CMC	33	6	Index MCP
11	10	Middle MCP	34	7	Index IP
12	11	Middle IP	35	8	Index TIP
13	12	Middle TIP	36	9	Middle CMC
14	13	Ring CMC	37	10	Middle MCP
15	14	Ring MCP	38	11	Middle IP
16	15	Ring IP	39	12	Middle TIP
17	16	Ring TIP	40	13	Ring CMC
18	17	Right heel	41	14	Ring MCP
19	18	Ring CMC	42	15	Ring IP
20	19	Ring MCP	43	16	Ring TIP
21	20	Ring IP	44	17	Right heel
	Body Pose No	Body Pose Name	45	18	Ring CMC
22	0	Nose	46	19	Ring MCP
23	11	Lef shoulder	47	20	Ring IP
24	13	Left elbow			

and improves the effectiveness of overall skeletal information structures. Directly affecting the movement of a joint skeleton is the main purpose of the motion calculation. There are 21 total landmarks in the joint skeleton dataset, and each landmark consists of x, y and z three-dimension coordinates; we captured motion for each coordinate separately [18]. The difference in the coordinated point between the consecutive frame joint positions is the concept of this calculation. The motion calculation procedure of the proposed method is visualized in Equation 1 and Figure 5.

$$\begin{aligned} Motion_X &= X_t - X_{t+1} \\ Motion_Y &= Y_t - Y_{t+1} \\ Motion_Z &= Z_t - Z_{t+1} \end{aligned} \tag{1}$$

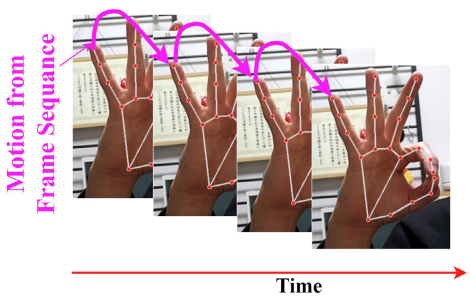


FIGURE 5. Visualize the motion calculation scenario.

C. GRAPH CONVOLUTION NETWORK (GCN)

GCN is a spatial type of neural network mainly designed to perform graph-structured data. In the study, we are working with the skeleton dataset, which mainly represents the human hand and body movement, which we represented with a graph shown in Figure 3. In the graph, we considered each joint a node and the connection between the two nodes as edges. The GCN model captured and then processed the skeleton graph as structural information based on the relationship among the nodes. In addition, the proposed graph structure captured the spatial relationship and dependencies among the joints.

D. GRAPH AND GRAPH CONVOLUTION CONSTRUCTION

In the study, we constructed a spatial-temporal graph to construct a hierarchical representation of the skeleton information extracted from the hand, body and face gestures [16], GSTCAN [17]. The constructed undirected graph from the skeleton can be expressed as the following Equation 2:

$$G = (V, E) \tag{2}$$

where the graph is constructed on the sequence of T frames and all N number of joints, our study has 32 frames and 47 skeleton joints based on the intra and inter-frame connection. Here V is mainly represented by the node-set number, which can be as $V = \{v_{(t,i)} \mid t = 1, \ldots, T, i = 1, \ldots, N\}$ by considering all the joints in each sequence. where $v_{(t,i)}$ represents the elements of the set V. The indices t and i range from 1 to T and N respectively. The convolutional graph here is constructed in the spatial and temporal domains. In the spatial domain, each joint of the human body is denoted as the vertex, and the spatial edge represents the natural connection of the human body. On the other hand, in the temporal domain, corresponding joints between the consecutive frames are considered a temporal edge [14], [18]. Multiple layers of the spatial and temporal GCN operation are employed here to predict the diverse action categories for extracting effective features. The GCN for the spatial domain for the specific vertex set v_i can be written as the following formula Equation.3:

$$F_{out}(v_i) = \sum_{(v_j) \in B_i) \frac{1}{Z_{ij}}} f_{in}(v_j) \cdot w(l_i(v_j)) \tag{3}$$

Here, the vertex of the graph is represented by v, the sampling area of the convolution for specific node v_i represented by B_i; it mainly denoted the distance between neighbour vertex v_i to v_j. W-like represents the waiting matrix generated function as an original convolutional layer, which generates the weight vector depending on the user's input terms.

E. FEATURE MAP OF THE GCN

The feature map of the GCN can be represented by the $C \times T \times N$ dimension, where $C, T, andN$ represent the number of channels, temporal length, and number of vertexes,

respectively. We can write the transformation formulas for implementing the proposed model according to the following Equation 4.

$$F_{GCN} = \sum_{k=1}^{K_v} W_k (f_{in} A_k) \otimes M_k \tag{4}$$

where F_{GCN} represents the output, $\sum$ denotes the summation, k is the index variable ranging from 1 to K_v, W_k represents a weight matrix, f_{in} is the input feature, A_k is a matrix, $\otimes$ denotes element-wise multiplication, and M_k is another matrix. Here, the kernel size of the spatial dimension is denoted by K_v. $A_k = \Lambda_k^{-\frac{1}{2}} \bar{A}_k \Lambda_k^{-\frac{1}{2}}$ here $N \times N$ adjacency matrix are represented by A_k and elements fo the $\bar{A}_k^{ij}$ represents the vertex v_j in the subject S_{jk} of the vertex v_i. It is mainly used to generate the connected vertexes for a particular subset from the input features by following the associated weight vector. After normalising the diagonal matrix, it can produce the following features $A_k^{ii} = \sum_j A_k^{-ij} + \alpha$ where α is employed to discard the empty row by setting value 0.001. The 1×1 convolution operation generated the weight vector W_k, which generated $C_{\text{out}} \cdot C_{\text{in}} \cdot 1 \cdot 1$. M_k is represented by the attention map generated $N \times N$ dimension and indicates the importance of various vertex and dot products denoted by $\odot$. In the same way, we applied the GCN for the temporal domain; also, it is easy to compare the spatial domain due to the number of neighbours for each vertex being fixed, which is 2, which indicates the previous and next consecutive frames. For the temporal domain, there needs $K_t \cdot 1$ convolution operation on the output of the feature map where k represents the kernel size for t time frames [14], [18].

F. CHANNEL ATTENTIOON

We applied the channel attention mechanism on the output of GCN to enhance the representation of features with the channel of a graph. Emphasizing the important feature suppresses the less important feature aiming to improve the gen-realization and discriminative power of the proposed system. In the previous GCN layer, we designed for operating the input skeleton data directly on the graph-structured data and computed the convolution operation for aggregating information from neighbouring nodes. The GCN generated the feature maps where each channel has a corresponding feature map. To refine the GCN feature, we employed channel attention, which explicitly enhances the interdependencies between the channels and dynamically adjusts their importance. By learning channel-wise weights, it can determine the contribution of each channel to the final representation. In the study, the channel attention models took the extracted feature maps, ran that feature throughout the global average pooling, and generated the output for each channel. After that, each channel is run through a couple of the fully connected layers, the batch normalization layer and coupled with the ReLU activation to produce the positive or *0* values. The powerful feature vector can be generated by multiplying the output of the activation function with the GCN features. In more explanation, the channel attention module produced a positive value for the potential features and *0* for the less effective features. After the multiplication operation, the important feature is selected from the GCN feature by converting the unimportant feature to zero [18]. Figure 6 demonstrates the structure of the channel attention used in the study where global average pooling took input from the N channels. We used a dense layer as size $N/8$ based and then passed it through a batch normalization layer to overcome the internal covariate shift problems and prevent the gradient from being too small. After using the ReLU activation layer, we used another fully connected layer whose size is N and fed into another ReLU activation. We focused on the ReLU activation because it has lower computational complexity than the sigmoid function.

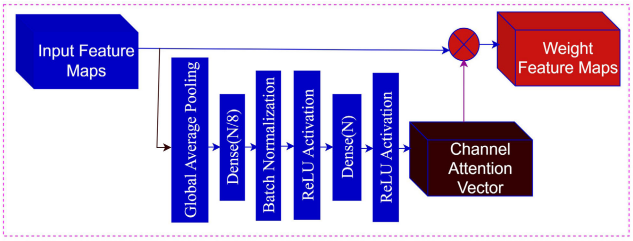


FIGURE 6. Architecture of the channel attention.

G. GENERAL-CNN BLOCK

In the stage, we employed a classification module where we included a coupled of the ReLU activation, Coupled of fully connected layer, layer normalization, dropout layer and the averaging pooling layer. Figure 7 demonstrates the classification module architecture.

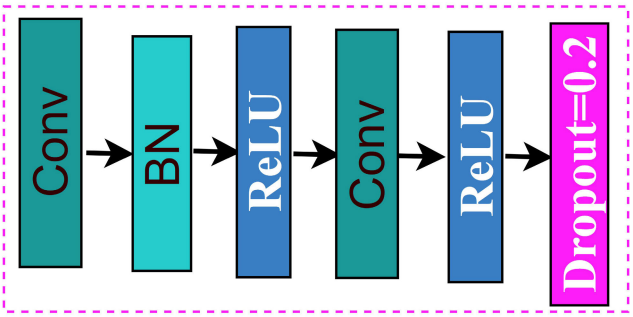


FIGURE 7. General CNN architecture.

H. CLASSIFICATION MODULE

In the stage, we employed a classification module where we included a coupled of the ReLU activation, Coupled of fully connected layers, layer normalization and dropout layer [4]. Figure 8 demonstrated the classification module architecture.

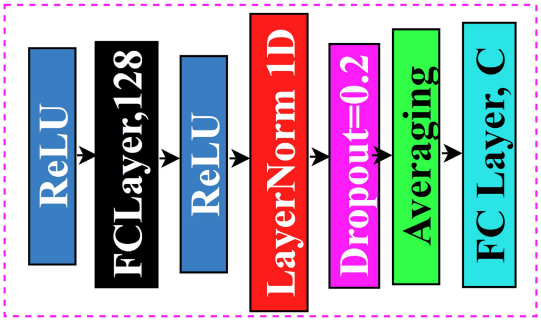


FIGURE 8. Classification module.

V. EXPERIMENTAL EVALUATION

We experimented with the two large Korean sign language datasets to test the superiority and effectiveness of the proposed study. To do this, firstly, we visualize the dataset with performance accuracy, and next, we demonstrate a comparison table with the state-of-the-art comparison. We divided the dataset into training and testing parts by following the approach of the previous paper. The gesture of the subject will be in the trained dataset, and the gesture subject will not be in the trained dataset.

A. ENVIRONMENTAL SETTING

To split the dataset into training and testing sets, we followed 7:3 rules where 70% is considered a training dataset, and 30% is a testing dataset. We used a Python environment and various Pytorch modules to implement the system. For the learning rate, we used a.001 value; we divided the 128 as batch size. We implemented the system in the Geforce RTX 4090 GPU machine, which consists of Linux and CUDA version 12.1, containing 64 GB RAM. We used Adam optimizer [58] with the Geforce RTX 4090 GPU and 100 epochs for the model run.

B. ABLATION STUDY

In our ablation study, we proposed a model structured with Graph Convolutional Network(GCN) modules spread across joint and joint motion streams with spatial-temporal contextual information enhancement. Within these, we integrated a combination of joint-based features and joint stream-based features supplemented by several Channel attention and General CNN modules. Building upon the foundational insights from prior studies, specifically [56] and [57], we gleaned a better understanding of the configuration of GCN and deep learning models. In our research, aiming for computational efficiency, we assessed configurations with a single GCN for spatial attention, accompanied by a Channel Attention and an NN module, as demonstrated in the ''Joint Stream'' and ''Joint Motion Stream'' models. Both models registered an accuracy of 96.22% on the KSL-77 dataset and 98.00% on the Proposed KSL-20 dataset. However, our standout result was achieved with the ''Proposed Model''. Here, we utilized 2 GCNs, supplemented by 2 Channel Attentions and 2 NN modules, in a parallel architecture emphasizing spatial and temporal contextual enhancements. This model yielded an impressive accuracy of 99.86% on the KSL-77 dataset and a perfect score of 100.00% on the Proposed KSL-20 dataset, marking a distinct edge over the other configurations. Another noteworthy configuration is the ''Two Stream Only GCN'' model, which utilizes 2 GCNs without any Channel Attention or CNN module. It achieved an accuracy of 99.90% on the KSL-77 dataset and 99.00% on the Proposed KSL-20 dataset.

TABLE 2. Strategic ablation study highlighting variations in GCN, channel attention and general CNN module counts aligned with spatial and temporal feature enhancement.

Method Name	No of GCN	No of Channel Attention	No of CNN Module	Accuracy KSL-77 Dataset [%]	Accuracy Proposed KSL-20 Dataset[%]
Joint Stream	1	1	1	96.22	98.00
Joint Motion Stream	1	1	1	96.22	98.00
Two Stream Only GCN	2	0	0	99.90	99.00
Proposed Model	2	2	2	99.86	100.00

These results unequivocally underscore the superior performance of our proposed configuration compared to other configuration architectures.

C. PERFORMANCE WITH THE BENCHMARK KSL DATASET

The study used two benchmark KSL datasets to evaluate the model: the publicly available large-scale KSL77 dataset and our proposed dataset. The KSL77 is considered one of the most challenging datasets in the KSL recognition domain. Classwise performance accuracy, precision, recall and f1 score are demonstrated in Table 3, and we showed here the performance matrix for the first 20 classes. We can see that the proposed model achieved good performance for all the labels equally. More than 50% classes achieved 100.00% accuracy, around 40% classes achieved more than 99.00% accuracy, and the rest of the classes produced more than 98.00% accuracy. Also reported the performance accuracy with the proposed KSL-20 dataset on the right side of the table, where our model generated more than 99.00% performance accuracy for all the classes.

The average of the performance matrix of all the classes is demonstrated in Table 4 for both datasets. We can see that our model produced the This demonstrated that the proposed model achieved 99.87%,99.87%,99.87%, and 99.87% for precision, recall, f1-score and accuracy, respectively, for benchmark KSL-77 datasets. In the same way, the next two visualized the proposed KSL-20 dataset; it produced the 100.00%, 100.00%, 100.00% and 100.00% for the precision, recall, f1-score and performance accuracy average.

D. COMPARISON OF THE STATE OF THE ART METHOD FOR KSL DATASET

We have included the state-of-the-art comparison for both datasets to prove the superiority of the proposed model. State-of-the-art comparisons for the existing KSL77 benchmark

TABLE 3. Precision, Recall and F1-Score for the KSL-77 and prposed KSL-20 dataset the first 20 classes.

Class no	KSL-77 Dataset			Proposed KSL-20		
	prec.	recall	f1-score	prec.	recall	f1-score
0	100.00	98.86	99.43	100.00	100.00	100.00
1	99.52	100.00	99.76	100.00	100.00	100.00
2	100.00	99.14	99.57	100.00	100.00	100.00
3	100.00	100.00	100.00	100.00	100.00	100.00
4	99.60	99.60	99.60	100.00	100.00	100.00
5	99.55	99.55	99.55	100.00	100.00	100.00
6	99.51	100.00	99.76	100.00	100.00	100.00
7	100.00	100.00	100.00	100.00	100.00	100.00
8	100.00	100.00	100.00	100.00	100.00	100.00
9	100.00	100.00	100.00	100.00	100.00	100.00
10	100.00	99.43	99.71	100.00	100.00	100.00
11	– –	– –	– –	100.00	100.00	100.00
12	100.00	100.00	100.00	100.00	100.00	100.00
13	100.00	100.00	100.00	100.00	100.00	100.00
14	100.00	100.00	100.00	100.00	100.00	100.00
15	100.00	100.00	100.00	100.00	100.00	100.00
16	99.57	100.00	99.79	100.00	100.00	100.00
17	100.00	99.13	99.57	100.00	100.00	100.00
18	– –	– –	– –	100.00	100.00	100.00
19	99.60	100.00	99.80	100.00	100.00	100.00

TABLE 4. Average Precision, Recall, F1-Score and Performance Accuracy for the two dataset.

Dataset Name	Av. Precision	Av. Recall	Av. F1-Score	Performance Accuracy[%]
KSL-77	99.87	99.87	99.87	99.87
Proposed Dataset	100.00	100.00	100.00	100.00

dataset are shown in Table 5. We can see that the existing model generated 79.80% and 93.00% accuracy, whereas the proposed model produced 99.87% performance accuracy, which is 6% more than the existing method. The existing method used a deep learning-based CNN model and reported 79.00% accuracy [40]. To improve the performance, shin et al. applied the multi-head-attention model to recognize the image pixel-based KSL dataset and achieved 93.00% accuracy [5]. They mainly focused on the various steps, including the grain module and parallel CNN with the multi-head attention model. Finally, they concatenated the features and used a classification module for the recognition. The proposed model produced 99.87% accuracy for the same dataset using a skeleton hand pose-based KSL dataset. We also experimented with the latest Graph-CNN-based state-of-the-art model, including Actional-structural graph convolutional networks (ASGCN) [17], graph-based spatial-temporal convolutional and attention neural network (DSTCAN) [18], and Spatial, temporal graph convolutional networks (DGSTA) [16], and we got 4.00%, 95.38% and 99.56% accuracy.

Table 6 demonstrates the comparison performance accuracy of the proposed model with the state-of-the-art comparison of the proposed model for the proposed KSL-20 dataset. The existing method generated 98.00% accuracy with the multi-stage attention-based model [56]. On the other hand,

TABLE 5. Comparison for the KSL-77 with the state-of-the-art model.

Name of the Dataset	Data Type	Name of the Model	Performance Accuracy[%]
KSL-77	Skeleton	ASGCN [17]	40.00
KSL-77	RGB Image	TSN [40]	79.80
KSL-77	RGB Image	Deep Learning [6]	93.00
KSL-77	join skeleton	GSTCAN [18]	95.38
KSL-77	join skeleton	DGSTA [16]	99.56
KSL-77	Joint Skeleton	Proposed Model	99.87

TABLE 6. Comparison for the proposed KSL-20 dataset with the state-of-the-art model.

Name of the Dataset	Data Type	Name of the Model	Performance Accuracy[%]
KSL-20	RGB Image	Shin et al [88]	98.00
KSL-20	Joint Skeleton	Proposed Model	100.00

the proposed model produced 100.00% accuracy, which is much more than the previous system.

E. DISCUSSION

In the study, we proposed a GCN with spatial-temporal attention and general neural recognition to recognize the KSL alphabet and world. In the study, firstly, we employed the Graph convolution to the skeleton dataset to convert the skeleton data into a graph structure. Then, we employed channel attention to produce the channel's effective feature. Then, we used adaptive graph CNN to extract features, and finally, we applied a classification module for the classification. Table 3 and Table 4 demonstrated the class-wise precision, recall, f1-score and performance accuracy and an average for all the classes, respectively. According to Table 5 and Table 6 performance, we can say that our model is much superior to the state-of-the-art method in the KSL recognition research domain. We believe that this study will upgrade the current KSL research situation and will be considered a novel method in this domain. Indeed, one of the key strengths of our vision-based Korean Sign Language recognition system lies in its accessibility. As a device-agnostic solution, our system can be seamlessly integrated into any device equipped with a camera. This versatility empowers a wide range of users to benefit from our application, including those who rely on smartphones, tablets, laptops, or even desktop computers. Whether it's a mobile device used on the go or a computer in a stationary setting, our system ensures that users have the freedom to communicate through sign language conveniently. This inclusivity aligns with our commitment to making communication barrier-free for all members of the community, including those who are hearing impaired or hard of hearing.

VI. CONCLUSION

In this study, we have pioneered the development of an advanced skeleton-based video classification system

featuring a two-stream attention-based neural network architecture. Each stream seamlessly integrates a Graph Convolutional Network (GCN) and an attention-driven neural framework, resulting in a robust and effective model for dynamic Korean Sign Language (KSL) recognition. In the first stream, we harnessed the complete body joint skeleton and meticulously processed it through a GCN, leading to the generation of robust graph-based features. Further enhancement of these features was achieved by incorporating a channel attention module, and their temporal context was enriched through the utilization of a general CNN. The second stream, mirroring the first, focused on capturing joint motion information. The fusion of features from both streams culminated in a powerful classification module, enabling precise sign language recognition.

Our model demonstrated exceptional performance in extensive evaluations, surpassing existing state-of-the-art models. This impressive achievement underscores the efficacy and superiority of our proposed model in the field of KSL recognition. Our model's adaptability, combined with its potential for training on various sign language datasets, opens doors to broader applications, benefiting not only the Korean hearing impaired community but also potentially extending its reach to other linguistic communities. In our future endeavours, we are committed to further optimizing our model. This includes expanding the dataset to encompass more sign words, selecting the optimal number of joints, enhancing feature extraction techniques, and leveraging advanced graph techniques to achieve even higher levels of performance. These ongoing efforts will not only improve the accuracy and efficiency of our model but also enhance its applicability in real-world scenarios. We believe these enhancements will solidify the position of our model as a promising solution for accessible communication through sign language recognition.

REFERENCES

[1] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[2] A. S. M. Miah, M. A. M. Hasan, J. Shin, Y. Okuyama, and Y. Tomioka, "Multistage spatial attention-based neural network for hand gesture recognition," *Computers*, vol. 12, no. 1, p. 13, Jan. 2023.

[3] A. S. M. Miah, J. Shin, M. A. M. Hasan, M. A. Rahim, and Y. Okuyama, "Rotation, translation and scale invariant sign word recognition using deep learning," *Comput. Syst. Sci. Eng.*, vol. 44, no. 3, pp. 2521–2536, 2023.

[4] A. S. M. Miah, J. Shin, M. A. M. Hasan, and M. A. Rahim, "BenSignNet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network," *Appl. Sci.*, vol. 12, no. 8, p. 3933, Apr. 2022.

[5] J. Shin, A. S. Musa Miah, M. A. M. Hasan, K. Hirooka, K. Suzuki, H.-S. Lee, and S.-W. Jang, "Korean sign language recognition using transformer-based deep neural network," *Appl. Sci.*, vol. 13, no. 5, p. 3029, Feb. 2023.

[6] H. Shin, W. J. Kim, and K.-A. Jang, "Korean sign language recognition based on image and convolution neural network," in *Proc. 2nd Int. Conf. Image Graph. Process.*, Feb. 2019, pp. 52–55.

[7] A. S. M. Miah, M. A. Rahim, and J. Shin, "Motor-imagery classification using Riemannian geometry with median absolute deviation," *Electronics*, vol. 9, no. 10, p. 1584, Sep. 2020.

[8] A. S. M. Miah, J. Shin, M. M. Islam, Abdullah, and M. K. I. Molla, "Natural human emotion recognition based on various mixed reality(MR) games and electroencephalography (EEG) signals," in *Proc. IEEE 5th Eurasian Conf. Educ. Innov. (ECEI)*, Feb. 2022, pp. 408–411.

[9] M. A. Rahim, A. S. M. Miah, A. Sayeed, and J. Shin, "Hand gesture recognition based on optimal segmentation in human-computer interaction," in *Proc. 3rd IEEE Int. Conf. Knowl. Innov. Invention (ICKII)*, Aug. 2020, pp. 163–166.

[10] A. S. M. Miah, M. A. Mouly, C. Debnath, J. Shin, and S. M. S. Bari, "Event-related potential classification based on EEG data using xDWAN with MDM and KNN," in *Proc. Int. Conf. Comput. Sci., Commun. Secur.* Cham, Switzerland: Springer, 2021, pp. 112–126.

[11] J.-S. Kim, W. Jang, and Z. Bien, "A dynamic gesture recognition system for the Korean sign language (KSL)," *IEEE Trans. Syst., Man Cybern., B, Cybern.*, vol. 26, no. 2, pp. 354–359, Apr. 1996.

[12] S. Yang, S. Jung, H. Kang, and C. Kim, "The Korean sign language dataset for action recognition," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2020, pp. 532–542.

[13] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Appl. Sci.*, vol. 9, no. 13, p. 2683, Jul. 2019.

[14] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.

[15] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, "Hand gesture recognition using deep feature fusion network based on wearable sensors," *IEEE Sensors J.*, vol. 21, no. 1, pp. 539–547, Jan. 2021.

[16] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 7444–7452.

[17] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3590–3598.

[18] R. Egawa, A. S. M. Miah, K. Hirooka, Y. Tomioka, and J. Shin, "Dynamic fall detection using graph-based spatial temporal convolution and attention network," *Electronics*, vol. 12, no. 15, p. 3234, Jul. 2023.

[19] S. C. W. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 873–891, Jun. 2005.

[20] D. M. Madhiarasan and P. P. P. Roy, "A comprehensive review of sign language recognition: Different types, modalities, and datasets," 2022, *arXiv:2204.03328*.

[21] J.-S. Kim, G. Park, Z. Bien, W. Jang, and S. Kim, "A dynamic pattern recognition system for the Korean sign language (KSL)," in *Proc. Asian Control Conf. (ASCC)*, Jul. 1994, pp. 713–716.

[22] C.-S. Lee, Z. Bien, G.-T. Park, W. Jang, J.-S. Kim, and S.-K. Kim, "Real-time recognition system of Korean sign language based on elementary components," in *Proc. 6th Int. Fuzzy Syst. Conf.*, vol. 3, 1997, pp. 1463–1468.

[23] M. M. H. Joy, M. Hasan, A. S. M. Miah, A. Ahmed, S. A. Tohfa, M. F. I. Bhuaiyan, A. Zannat, and M. M. Rashid, "Multiclass MI-task classification using logistic regression and filter bank common spatial patterns," in *Proc. Int. Conf. Comput. Sci., Commun. Secur.* Singapore: Springer, 2020, pp. 160–170.

[24] T. Zobaed, S. R. A. Ahmed, A. S. M. Miah, S. M. Binta, M. R. A. Ahmed, and M. Rashid, "Real time sleep onset detection from single channel EEG signal using block sample entropy," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 928, Jul. 2020, Art. no. 032021.

[25] M. H. Kabir, S. Mahmood, A. Al Shiam, A. S. M. Miah, J. Shin, and M. K. I. Molla, "Investigating feature selection techniques to enhance the performance of EEG-based motor imagery tasks classification," *Mathematics*, vol. 11, no. 8, p. 1921, Apr. 2023.

[26] A. S. M. Miah, J. Shin, M. A. M. Hasan, M. K. I. Molla, Y. Okuyama, and Y. Tomioka, "Movie oriented positive negative emotion classification from EEG signal using wavelet transformation and machine learning approaches," in *Proc. IEEE 15th Int. Symp. Embedded Multicore/Many-Core Systems Chip (MCSoC)*, Dec. 2022, pp. 26–31.

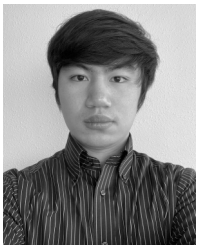
[27] K. A. Kibria, A. S. Noman, M. A. Hossain, M. S. I. Bulbul, M. M. Rashid, and A. S. M. Miah, "Creation of a cost-efficient and effective personal assistant robot using Arduino & machine learning algorithm," in *Proc. IEEE Region 10th Symp. (TENSYMP)*, Jun. 2020, pp. 477–482.

[28] A. S. M. Miah, J. Shin, M. A. M. Hasan, Y. Fujimoto, and A. Nobuyoshi, "Skeleton-based hand gesture recognition using geometric features and spatio-temporal deep learning approach," in *Proc. 11th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Sep. 2023, pp. 1–6.
[29] M. A. Uddin and S. A. Chowdhury, "Hand sign language recognition for Bangla alphabet using support vector machine," in *Proc. Int. Conf. Innov. Sci., Eng. Technol. (ICISET)*, Oct. 2016, pp. 1–4.
[30] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos, "Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition," in *Proc. CVPR WORKSHOPS*, Jun. 2011, pp. 1–6.
[31] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2200–2207.
[32] S. G. M. Almeida, F. G. Guimarães, and J. A. Ramírez, "Feature extraction in Brazilian sign language recognition based on phonological structure and using RGB-D sensors," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7259–7271, Nov. 2014.
[33] R. Fatmi, S. Rashad, and R. Integlia, "Comparing ANN, SVM, and HMM based machine learning methods for American sign language recognition using wearable motion sensors," in *Proc. IEEE 9th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2019, pp. 0290–0297.
[34] Y. Na, H. Yang, and J. Woo, "Classification of the Korean sign language alphabet using an accelerometer with a support vector machine," *J. Sensors*, vol. 2021, pp. 1–10, Aug. 2021.
[35] A. Y. Dawod and N. Chakpitak, "Novel technique for isolated sign language based on fingerspelling recognition," in *Proc. 13th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA)*, Aug. 2019, pp. 1–8.
[36] V. T. Hoang, "HGM-4: A new multi-cameras dataset for hand gesture recognition," *Data Brief*, vol. 30, Jun. 2020, Art. no. 105676.
[37] C. Chansri and J. Srinonchat, "Hand gesture recognition for Thai sign language in complex background using fusion of depth and color video," *Proc. Comput. Sci.*, vol. 86, pp. 257–260, Dec. 2016.
[38] S. P. Y. Jane and S. Sasidhar, "Sign language interpreter: Classification of forearm EMG and IMU signals for signing exact English," in *Proc. IEEE 14th Int. Conf. Control Autom. (ICCA)*, Jun. 2018, pp. 947–952.
[39] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche, "Hand gesture recognition for sign language using 3DCNN," *IEEE Access*, vol. 8, pp. 79491–79509, 2020.
[40] 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.
[41] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Syst. Appl.*, vol. 150, Jul. 2020, Art. no. 113336.
[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[44] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.
[45] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[48] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[50] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.
[51] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[52] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.
[53] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," 2021, *arXiv:2102.10882*.
[54] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
[55] Y. Ji, S. Kim, and K.-B. Lee, "Sign language learning system with image sampling and convolutional neural network," in *Proc. 1st IEEE Int. Conf. Robotic Comput. (IRC)*, Apr. 2017, pp. 371–375.
[56] A. S. M. Miah, M. A. M. Hasan, S.-W. Jang, H.-S. Lee, and J. Shin, "Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition," *Electronics*, vol. 12, no. 13, p. 2841, Jun. 2023.
[57] A. S. M. Miah, M. A. M. Hasan, and J. Shin, "Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model," *IEEE Access*, vol. 11, pp. 4703–4716, 2023.
[58] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proc. ICLR*, 2016, pp. 1–4.

JUNGPIL SHIN (Senior Member, IEEE) received the B.Sc. degree in computer science and statistics and the M.Sc. degree in computer science from Pusan National University, South Korea, in 1990 and 1994, respectively, and the Ph.D. degree in computer science and communication engineering from Kyushu University, Japan, in 1999, under a scholarship from the Japanese government (MEXT). He was an Associate Professor, a Senior Associate Professor, and a Full Professor with the School of Computer Science and Engineering, The University of Aizu, Japan, in 1999, 2004, and 2019, respectively. He has coauthored more than 350 published papers for widely cited journals and conferences. His research interests include pattern recognition, image processing, computer vision, machine learning, human–computer interaction, non-touch interfaces, human gesture recognition, automatic control, Parkinsons disease diagnosis, ADHD diagnosis, user authentication, machine intelligence, bioinformatics, and handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He serves as a reviewer for several major IEEE and SCI journals. He served as the program chair and a program committee member for numerous international conferences. He serves as an Editor for IEEE journals, Springer, Sage, Taylor and Francis, MDPI *Sensors* and *Electronics*, and Tech Science. He serves as an Editorial Board Member for *Scientific Reports*.

ABU SALEH MUSA MIAH received the B.Sc.Eng. and M.Sc.Eng. degrees in computer science and engineering from the Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh, in 2014 and 2015, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, The University of Aizu, Japan. He became a Lecturer and an Assistant Professor with the Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh, in 2018 and 2021, respectively. He received a scholarship from the Japanese government (MEXT) for the Ph.D. study. He has authored or coauthored more than 20 publications and has published in widely cited journals and conferences. His research interests include CS, ML, DL, HCI, BCI, and neurological disorder detection.

KOTA SUZUKI is currently pursuing the bachelor's degree in computer science and engineering with The University of Aizu (UoA), Japan. He joined the Pattern Processing Laboratory, UoA, in April 2023, under the direct supervision of Dr. Jungpil Shin. He is currently working on fire detection and human activity recognition. His research interests include computer vision, pattern recognition, and deep learning.



KOKI HIROOKA was born in Aizumisato, Fukushima, Japan. He received the bachelor's degree in computer science and engineering from The University of Aizu (UoA), Japan, in March 2022, where he is currently pursuing the master's degree. He joined the Pattern Processing Laboratory, UoA, in April 2021, under the supervision of Dr. Jungpil Shin. He is currently working on human activity recognition, human gesture recognition, Parkinsons disease diagnosis, and ADHD diagnosis. His research interests include computer vision, pattern recognition, and deep learning.

MD. AL MEHEDI HASAN (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science and engineering from the Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh. He became a Lecturer, an Assistant Professor, an Associate Professor and a Professor with the Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology (RUET), Rajshahi. Recently, he completed a postdoctoral research with the School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan. He has coauthored more than 130 publications and has published in widely cited journals and conferences. His research interests include machine learning, deep learning, bioinformatics, health informatics, computer vision, probabilistic and statistical inference, medical image processing, sensor-based data analysis, human–computer interaction, operating systems, computer networks, and security.