IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 December 2023, accepted 9 January 2024, date of publication 16 January 2024, date of current version 24 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354775

RESEARCH ARTICLE

Joint Motion Affinity Maps (JMAM) and Their Impact on Deep Learning Models for 3D Sign Language Recognition

P. V. V. KISHORE[1], (Senior Member, IEEE), D. ANIL KUMAR[2], (Member, IEEE), RAMA CHAITHANYA TANGUTURI[3], (Senior Member, IEEE), K. SRINIVASARAO[4], P. PRAVEEN KUMAR[5], AND D. SRIHARI[6], (Member, IEEE)

[1]Department of Electronics and Communication Engineering, Biomechanics and Vision Computing Research Center, Koneru Lakshmaiah Education Foundation (Deemed–to–be–University), Guntur 522502, India
[2]Department of Electronics and Communication Engineering, PACE Institute of Technology and Sciences, Ongole 523272, India
[3]Department of Computer Science and Engineering, PACE Institute of Technology and Sciences, Ongole 523272, India
[4]Department of Electronics and Communication Engineering, Dhanekula Institute of Engineering and Technology, Vijayawada, Andhra Pradesh 521139, India
[5]Department of Information Technology, Vignan's Institute of Information Technology, Duvvada, Visakhapatnam 530049, India
[6]Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering and Technology, Chittoor 517127, India

Corresponding author: P. V. V. Kishore (pvvkishore@kluniversity.in)

ABSTRACT Previous works on 3D joint based feature representations of the human body as colour coded images (maps) were developed based on the joint positions, distances and angles or a combination of them for applications such as human action (sign language) recognition. These 3D joint maps have shown to singularly characterize both the spatial and temporal relationships between skeletal joints describing an action (sign). Consequently, the joint position and motion identification problem transformed into an image classification problem for 3D skeletal sign language (action) recognition. However, the previously proposed process of transforming 3D skeletal joints to colour coded maps has a negative proportionality component which resulted in a map with small pixel densities when the joint relationships are high. This drawback greatly impacts the learning of the classifiers to quantify the joint relationships within the colour coded maps. We hypothesized that a positive proportionality between joint motions and the corresponding maps would certainly improve classifiers performance. Hence, joint motion affinity maps(JMAM). These JMAMs use radial basis kernel on joint distances which assures a positive proportionality constant between joint motions and pixel densities of colour coded maps. To further improve the classification of 3D sign language, this work proposes congruent body part joints which results in motion directed JMAMs with maximally discriminating positive definite spatio temporal features. Finally, JMAMs are trained on the proposed multi-resolution convolutional neural network with spatial attention (MRCNNSA) architecture which produces an influencing result for the constructed 3D sign language data, KL3DISL. Consequently, online 3D datasets and standard deep learning models benchmark the proposed with respect to sign and action recognition. The results conclude that JMAMs with clustered joints characterize the subtle relationships which are otherwise difficult to be learned by a classifier.

INDEX TERMS 3D joint data processing, motion affinity maps, radial basis kernel, neural network with self-attention.

I. INTRODUCTION

Despite 3 decades of research on sign language recognition (SLR), quantifying subtle finger movements in video sequences remains a challenge. While hand gloves [1], [2], [3] have shown high transformation efficiency for SLR, their high cost make it difficult to reach open markets, and their service to hearing-impaired individuals remains unrealized. Furthermore, visual nature of sign language cannot be disregarded, which is why a significant portion

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

of SLR research has focused on 2D video data. Moreover, it capitalizes on the core components of vision such as spatial shapes and semantic movements of hands, fingers, face, head, and torso of the signer.

Overall, the 2D RGB video data [4], [5] for sign language (SL) has dominating advantages over its counterparts 1D and 3D such as more information, relatedness and economical. However, 2D data has always been demanding improved techniques in the areas of feature engineering [6], data representation [7] and classifier design [8]. On the contrary, significant impact has been shown by the application of deep neutral networks (DNN) to recognize sign language from 2D video data. Although the inference accuracy of DNNs on video SL data [9], [10] was found to be significant, their confidence to discriminate signs with similar movements and closely matched finger shapes is low. As a result, there are methods that tried to improve on this through deeper networks [11], [12], multi-modality training data [13], [14] and hybrid feature embeddings [15], [16]. More noticeable difference in accuracies were recorded when the training data consisted of multiple modalities such as RGB, depth and skeletal [17], [18]. Impressive results were detected with 3D skeletal data with sequence modeling deep networks such as long short-term memory (LSTMs) [19], [20] for both sign language and human action recognition applications. In spite of this being the best performing network, it fails to quantify the spatial joint relations within sign (action). The consequences of this on skeletal sign language methods were more than noticeable during testing where the trained network fails to converge efficiently towards the target labels.

On the other hand, the use of 3D joint-based feature representations has demonstrated its effectiveness in various fields of computer vision, including biomechanics, human-computer interactions and robotics. These representations capture spatial and temporal aspects of movements in 3D space that are recorded using motion capture technology. The challenges such as background variations, lighting conditions and object occlusions that are part of RGB video based SLR are rendered ineffective in 3D motion capture system. However, it is impractical to use motion capture technology for real-time 3D SLR applications. Therefore, a more practical approach is to apply 3D pose estimation to RGB video data and construct a spatiotemporal feature map of a sign. Then, compare the estimated pose map to quantify the 3D pose map representations from the motion capture system for recognition. The major advantage is the joint information coverage in 3D pose estimations using the 3D motion capture data, which makes it a reliable system for real time operation. Nevertheless, this real-time system requires dedicated hardware to avoid latency when detecting the correct motion-captured pose from the reconstructed estimated 3D pose from 2D video data.

To accommodate both spatial and temporal characteristics simultaneously, the 3D data is represented as a colour coded image. The 3D joint features such as distance and angles were converted to RGB coded images known as

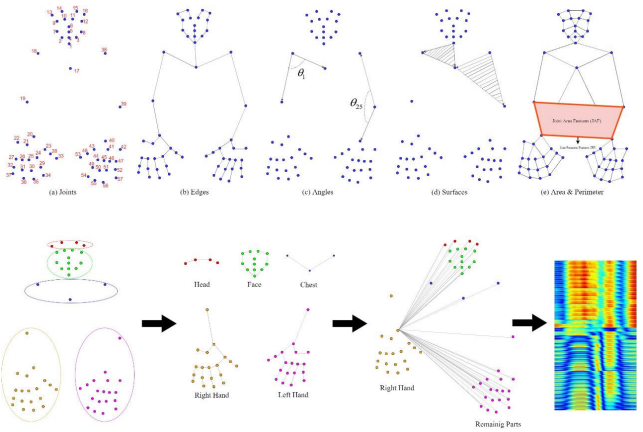

FIGURE 1. A comparison of colour coded feature maps from 3D joint positional data across the frames. The last column shows jg2gMDMs that represents high pixel density for motion joints when compared to non-motion joints on the skeleton.

joint distance maps [21], [22] and joint angular maps respectively [18], [23]. Consequently, the colour maps trained on convolutional neutral networks (CNN) [24], [25] has showed better performance across all the sign language recognition methods. The biggest problem again lies in the poor representation of signs having more than 60% matched action content. For example, head, hair, forehead, face and mother, female, woman, girl. Until now, the colour coding is directly proportional to the observable joint variations such as positions, angles and their derivatives. These are distance, velocity, acceleration, angular velocity, momentum etc. These color-coded join feature maps show shoddy relationships between highly interconnected joints of a sign and vice versa. Consequently, the trained classifiers were unable to capture highly contributing joint relationships correctly to identify a particular sign. This is because of the lowest pixel values that are directly proportional to the highest joint relativity. The proposed Joint Motion Affinity Maps (JMAM) are based on an inverse relationship of joint features with the pixel representations. It has shown to improve the quality of color-coded features for classification and there by the recognition accuracy of skeleton based 3D sign language. Figure.1 shows a formal comparison between the colour coded feature maps from previous works and the proposed JMAM.

To improve the performance of SLR, this study proposes clustering of congruent body part joints. The top row of Figure.2 shows part-based models for skeleton-based human action recognition (HAR) models from the literature [26], [27] which are presented in Figures.2(a)-(e). In the bottom row of Figure.2, we present the division of joint groups on our 57 joint SL skeleton. The entire SL skeleton is divided into 5 parts as shown in 1^{st} sub-figure of 2^{nd} row in Figure.2. Subsequently, clustering results in joint group to group motion directed JMAMs (jg2gMDMs), which maximize the discriminative positive definite spatiotemporal features. The resulting jg2gMDMs, trained on any standard

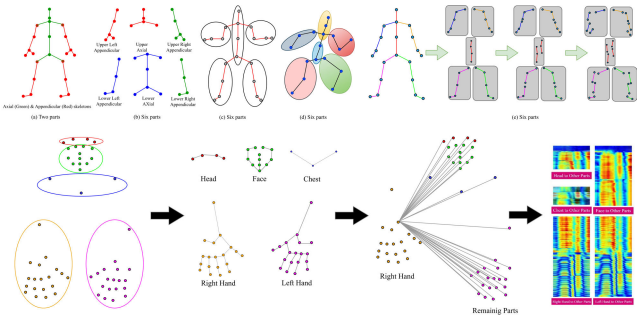

FIGURE 2. Part based joint clustering with different joint combinations for HAR. (a) Two parts clustering, (b) Six parts, (c) Five Parts, (d) Six Parts, (e) Multiple Groups and (f) Five part division of sign language skeleton and construction joint group to group movement affinity maps (jg2gMAMs).

deep learning architectures, have reduced false positives and true negatives that were common in SLR. Additionally, the proposed color-coding model has been adopted for validating benchmark action recognition datasets, such as NTU RGBD [28], KL3DYOGA [23], CMU [29] & HDM05 [30]. The results show that JMAMs and jg2gMDMs transform the subtle joint relationships between closely matched signs that are otherwise challenging for a classifier to learn. The classifier is a lightweight convolutional neural network with spatial and channel-based self-attention architecture across multi-resolution inputs, producing an influential result for color-coded 3D joint data.

The following are the major contributions of this paper:

1) A 3D skeletal sign language database comprising 500 signs is created using 10 subjects. Each subject performs each sign 5 times, resulting in a total of $500\times10\times5 = 25000$ signs.
2) The paper proposes the use of Joint Motion Affinity Maps (JMAMs) to emphasize the motion relativity between skeleton joints through color-coded embeddings.
3) The joint group to group motion directed JMAMs (jg2gMDMs) representing grouped motion information in colour-coded embedding maps.
4) Lightweight multi-resolution feature fusing convolutional neural network with spatial attention mechanism for classification.

The rest of the paper organizes into a literature review, methodology, experimentation, and conclusions in sections II, III, IV, and V respectively.

II. BACKGROUND

SLR has been studied using various data [4] sources, such as hand gloves (1D) [1], video cameras (2D) [5] Kinect or leap motion (3D) [17], and the high-priced motion capture technology employed in this work [18]. In our opinion, the best results can be obtained using 3D motion capture datasets. However, the most used data source is 2D video data. While Microsoft Kinect's 3D skeletal sign language data lacks finger movements, leap motion only captures hand movements. The 3D motion capture system provides a rich representation of the human skeleton with 57 joints in space eliminating 2D anomalies such as lighting, recording speed, camera angles, motion blurring, and occultations [31]. Although expensive, the 3D motion capture signs exhibit a naturalistic resemblance to real-time human actions with superior representations compared to other sources.

In comparison to 3D skeletal, the most used source for SL was 2D RGB video data [32]. A wide variety of algorithms were proposed in the last few decades for video preprocessing, feature extraction and recognition [32], [33]. Most of these algorithms solved some type of spatial, temporal or paired representation of video object data effectively as features. These features are further classified using all the traditional machine learning algorithms. The most popular classifier were, Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) [34]. With the advent of deep learning frameworks, the 2D video based SLR has become powerful with the option of feature learning rather than feature extraction. A large contingent of them are available for perusal [35]. The accuracies reported by these methods are not reproduceable or they simply fail to generalize on the video quality or the signer. This has motivated researchers towards higher dimensional data such as RGB D or 3D skeletal representations. Multi modal video sequences that are fed into multiple streams of a CNN are predominantly researched which have shown evidence of exceptional performances in real time for sign (action) recognition applications [36]. The recognition accuracies were better than the single modal datasets. However, the training requires higher computing powers. At the same time, working with different modalities such as skeletal and depth information has shown to enhance the performance of recognition algorithms in Indian and Russian sign language models [37], [38]. The use of generative AI models has improved the performance of the classifiers.

Human skeletal data is more robust than other modalities, such as RGB video and depth, due to its independence towards video backgrounds and human subject inconsistencies. This has made the 3D skeletal representation of human actions and activity as the preferred input modality for classification. The availability of inexpensive hardware sensors, such as Microsoft Kinect and Intel RealSense 3D capture system [39], [40], has further fuelled this trend. More expensive and accurate capture technology, such as multi-camera 3D mocap system [18], has also revolutionized human action recognition in the last ten years. While a multitude of action recognition algorithms has been proposed on these datasets [28], [29], [30], we will focus on reviewing works that use deep learning frameworks.

The ideal approach for recognizing actions based on 3D skeletal data is to combine joint action data with deep learning. Skeletal data is spatially relational, temporally compatible, and can form spatio temporal structures, making

it an ideal input modality for automated skeletal action recognition [41]. Early machine learning models focused on learning temporal patterns by extracting joint variations across frames [42] and characterizing them as time series representations [43]. However, these models learned temporal variations specific to a dataset and could not transfer the gained knowledge during testing with a different dataset. Recurrent ML models were found to perform poorly across datasets. To develop actionable intelligence across datasets, deep learning architectures were applied to skeletal action data. Previous deep learning models used in vision computing applications [23], [44] have shown impressive performances in decoding spatial and spatio-temporal patterns.

Deep learning methods such as Recurrent Neural Networks (RNNs) [45], Long Short Term Memory (LSTM) [46], Convolutional Neural Networks (CNN) [47], Recurrent CNN (RCNN) [48] and Graph Convolutional Networks (GCN) [49] have shown remarkable progress in human action recognition with skeletal datasets [23], [28], [30], [50]. RNNs [45] have been successful in characterizing naturally occurring skeletal joint temporal cues in human actions. The structure of RNNs enables them to identify joint patterns by establishing relationships between previous and present joint variations across action sequences. However, RNNs have limitations in processing long sequences due to the vanishing gradients problem [45]. To overcome this issue, memory cells have been incorporated into the current architecture of RNNs, resulting in upgrades such as LSTM and Gated Recurrent Units (GRU).

LSTMs have been primarily used for skeletal action recognition tasks in both unidirectional [51] and bidirectional modes [52]. Among the LSTM models, bidirectional LSTMs have shown to achieve higher recognition accuracies [52]. However, LSTMs are computationally intensive and can suffer from the vanishing gradient problem due to the tanh function. Independent recurrent neural networks have been proposed as an improved architecture to address these issues, allowing for longer and deeper architectures without the vanishing gradient problem [52]. However, recurrent models may lack the ability to capture spatial features that define joint relationships within a skeletal action frame. To address this, spatial temporal combination networks have been proposed, using CNNs [53], [54] to learn spatial joint features and then inputting the flattened features into LSTMs to determine temporal patterns in the extracted spatial contents. Despite achieving higher recognition accuracies CNN-LSTM models require significant computational power during training on large human action datasets [53], [54].

To overcome these network implications for action recognition, a rich spatio temporal feature representation is uncovered in the form of RGB color images. These RGB color maps characterize a particular skeletal action across a set of 3D video frames. Consequently, the proposed spatio temporal images are found to be independent of length of the video sequences as well as number of joints. These spatio temporal features express spatial relationships among joints within a 3D action frame as well as temporal changes between frames. The previously proposed spatio temporal features are Joint Positional Maps (JPM), Joint Distance Maps (JDM), Joint Angular Maps (JAM), Joint Angular Displacement Maps(JADM), Joint Velocity Maps (JVM), Joint Acceleration Maps (JaM), Joint Planar Maps (JpM), Joint Trajectory Maps (JTM) and Joint Quadrilateral Volume Maps (JQVM). The above spatio temporal feature maps are embedded with patterns that can be quantified using a deep CNN of any architecture. It has been shown that the deep CNNs had certainly enhanced the performance of the skeletal action recognition system on Kinect and mocap datasets.

III. MAPPING JOINTS TO COLOUR CODED IMAGES

Skeletal 3D sign language has been traditionally captured using sensors such as Kinect [17] which are unable to accurately capture overlapping joints. This results in lower classifier performance due to missing or overlapping joint features. To address this issue, we used motion capture technology build our 3D skeletal sign language datasets [18], [31]. This section describes the methodology for computing color-coded feature maps from the positional information of the 3D skeleton sign language actions. The first half of this section reviews previously used methods for computing color-coded features, while the second half discuss the limitations of existing mapping processes and introduces Joint Motion Affinity Maps (JMAM) as a solution.

A. ENCODING 3D JOINT POSITIONS INTO MAPS

The 3D skeleton joints, shown in Figure.2, are described with respect to the 3D world origin set during initialization. Each 3D joint J_p is represented as $(x_i, y_i, z_i) \in R^{3\times p}$ where $i = 1\ to\ p$ within a video frame $f = 0\ to\ F \in I^+$. In the past, it has been shown that working directly on $J'_p s$ as a time-series data and exploiting the recurrence phenomenon has resulted in a test accuracy of around 82%. Additionally, the recurrent network operated at optimal speeds during inference. However, for 3D sign language or actions with overlapping joint positions, the recurrence method has shown a test accuracy of less than 40%. This is due to the fact that in 3D SL, the hand joints are more likely to interact with any of the body parts, as compared to an action sequence such as walking or running. Moreover, the above time-series representation of $J'_p s$ across frames in SL will miss the spatial information that models the relation of moving hands with respect to the torso in defining a sign.

The above limitation has been addressed by using JMAM which convert the problem into an image classification task on a convolutional neural network. These maps represent both spatial and temporal joint movements using color-coded images. Previous works [18], [23] proposed various maps, such as JDMs, JAMs, JADMs, JVMs, JaMs and JamMs, to characterize joint-to-joint relationships. However, these maps impacted the classifier's inferencing accuracy to a certain extent. To compute JDMs on a J joint skeleton with F video frames consisting of J_{C_2} pairs, we compute the d^c

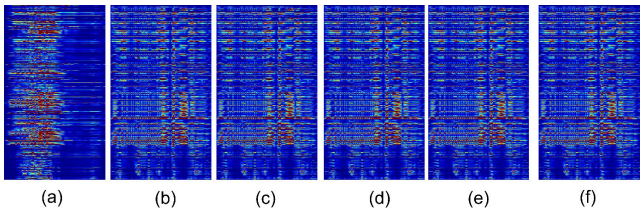

FIGURE 3. Spatiotemporal color-coded maps of 3D sign 'basketball'. (a) JDMs, (b) JAMs, (c) JADMs, (d) JVMs, (e) JaMs and (f) JamMs.

matrix given as:

$$d^c = Concat[d_x; d_y; d_z] \; \forall \; x, y, z, c \tag{1}$$

where, c is the class representative and $d_{(\blacksquare)}$ is the distance between unique joint pairs. The Euclidean distance function calculates the norm between i^{th} and j^{th} joints J_p in x, y, z planes separately with

$$d^{ij}_{(\bullet)} = \left\| p_i - p_j \right\|_2 \forall \, (x, y, z) \tag{2}$$

For $\frac{p(p-1)}{2}$ each of the unique joint positional pairs within a $f(x, y, z)$ frame in three dimensions, the $d^{ij}_f = \left[d^f_x; d^f_y; d^f_z \right] \in R^{\frac{p(p-1)}{2} \times 3}$ distance between them is calculated. Finally, the joint distance matrix F for all frames will have a dimensionality of $\frac{p(p-1)}{2} \times 3 \times F$. To construct JDMs from joint distance matrix, the three dimensions are color coded with JET colour coding as Red, Green and Blue planes of an image. Consequently, one JDM image characterizes a class label in the dataset which represents spatiotemporal relationships between joints in the entire 3D video sequence.

Similarly, the JAMs quantify the relationships between joints in a 3D sign video by computing the

$$\emptyset^f_{ij} = \cos^{-1} \left(\frac{\overrightarrow{p_{ik}} \times \overrightarrow{p_{kj}}}{\sqrt{\overrightarrow{p_{ik}}} \sqrt{\overrightarrow{p_{kj}}}} \right) \forall f \tag{3}$$

The joint angle $\emptyset^f_{ij}$ gives the inclination of positional vectors j and i with respect to joint k. In order to construct a RGB color coded map, the above angular information is mapped in 3D along all F frames into a $\frac{p(p-1)}{2} \times F$ image. Subsequently, the joint angular displacement maps (JADMs) are constructed by using

$$d^f_{\emptyset} = d^{ij}_f \times \cos \left(\emptyset^f_{ij} \right) \forall f = 1 \; to \; F \tag{4}$$

Contrastingly, the above distance maps were augmented in spatiotemporal representations by computing the joint distances and angles between frames to construct velocity(JVMs), acceleration (JaMs) and angular momentum maps (JamMs). Figure.3 shows all the maps for our 3D sign language dataset sign 'basketball'.

The horizontal axis represents the features and the vertical axis the frames in a 3D sign or action video sequence. These maps for all skeletal classes in the dataset are given as input to the convolutional neural networks with few trainable parameters. All these maps performed well with training accuracies ranging from 85 to 98%. However, the inferencing on these trained models have shown unsatisfactory results due to two major issues in the maps. The first one being the direct proportionality between the extracted features such as distance, angles or others and the pixel intensities. This resulted in low intensity pixels on the maps when the joint distances(angles) are decreasing and vice versa. Figure.3 clearly marks the regions as low pixel intensities with decreasing joint distances. This disadvantage gets further magnifies during the training process which gives more attention to retracting joint features than the contracting ones. This is the second major problem in the current generation of 3D skeletal sign language recognition models from feature maps.

Additionally, the models are unable to extract information in signs that have similar action content. For example, signs like "woman" and "mother", "head" and "hair", "food" and "hungry", "drink" and "tea" to name a few. Figure.4 shows the 3D signs and their color-coded maps from the features extracted previously. To overcome this problem, we explored the part based [55], [56] color-coding process where we divide the 3D skeleton into 5 parts. They are head, face, chest, Right hand and left hand as shown in Figure.2. To overcome all the above problems, we propose to instigate a positive proportionality between joint motions and the corresponding maps that would certainly improve classifiers performance. Hence, joint motion affinity maps (JMAM) are being proposed on full body and congruent body part joints which results in motion directed JMAMs and jg2gMDMs with maximally discriminating positive definite spatio temporal features respectively. The jg2gMDMs is a cluster of joints representing each body part and their motion characterization using a color-coded map. These maps has packed more local information between body joint groups that would facilitate more quick and computationally inexpensive models and training respectively. Finally, to improve the recognition on the 3D skeletal joint full and split color-coded features, we propose multi resolution convolutional neural network with spatial attention (MRCNNSA) architecture which produces an influencing result for 3D skeletal sign language data. The following subsection gives an elaborate discussion on the computation of joint movement affinity maps (JMAMs).

B. JOINT MOVEMENT AFFINITY MAPS (JMAMS)

The J_p 3D joints in the skeleton form $\frac{J_p(J_p-1)}{2}$ unique pairs. The Euclidean distance $d^{ij}_{(\bullet)}$ on these pairs in each frame f are computed as shown in Equation(1). The $(\bullet) \rightarrow (x, y, z)$. This results in $\left\{ d^{ij}_x, d^{ij}_y, d^{ij}_z \right\} \in R^{\frac{J_p(J_p-1)}{2} \times 1}$, which capture the spatial distribution of joints within the frame f. Enumerating on all the frames in a class label results in a feature matrix $\Psi_f = \left[d_x; d_y; d_z \right] \in R^{\frac{J_p(J_p-1)}{2} \times 3} \forall f = 1 \rightarrow F$. To transform these Euclidean distance features $\left\{ \Psi_f \right\}$ into a movement

affinity feature matrix $M(\bullet)$, we compute the following kernel function given as

$$M(\bullet) = \exp\left(-\frac{(\Psi_f - \Psi_{f+1})}{2\xi^2}\right) \forall (x, y, z, f) \quad (5)$$

Solving for Equation(5) with result in three movement affinity feature matrices in (x, y, z) directions between F frames. We have $M = [M(x), M(y), M(z)] \in R^{\frac{J_p(J_p-1)}{2} \times 3 \times F-1}$. These features represent spatial movements across the frames. Finally, the computed movement affinity feature matrices in (x, y, z) directions are JET color coded as Red, Green and Blue planes. The JET color coded M is called Joint Movement Affinity Map (JMAM). Each class of 3D joint data is represented by a single spatiotemporal color-coded JMAM. Figure.4 shows the JMAMs for 3D signs of our dataset. The figure shows key motion frames of 3D sign language data and their corresponding spatiotemporal maps. It is a comparison between the previously constructed feature maps versus the JMAMs from this work. As discussed, the previously constructed 3D joint features in space and motion, such as JDMs, JADMs, etc., are in direct proportion to the color-coded pixel intensities. This implies that the decreasing joint distances during a signing process are mapped into low pixel values and similarly, the increasing joint distances are mapped to the largest pixel values. As a result, the dynamic range of motion mapped pixels becomes wide enough to have significant impact on the classifier performance during training.

In contrast to the above, the JMAM features ensure a good dynamic range between pixels for decreasing as well as increasing 3D joint distances. By applying Equation(5), the pixel values are always oriented in the direction of increasing gradient. To elaborate, when the 3D joint distances are decreasing, the pixel values are increasing with a marginal positive gradient, and for increasing distances the pixel values will have a high positive gradient. This distinction can be observed in the most common signs shown in Figure.4. Even more important are the feature variations in 3D finger joints that characterize a sign. These intra hand finger joint feature variations are relatively small compared to the overall hand motions. This has indeed become a problem in the previously proposed maps where these small finger variations are mapped to lower pixel values and large hand motions into higher ones. This is contrary to the mapping procedure followed in Equation(5) for JMAMs. Which is evident visually through observation of maps in Figure.4.

Moreover, the use of JMAMs is also beneficial in mitigating the issue of large inter-subject variations by maintaining an acceptable margin of pixel dynamic range. We set the value of sigma to 1.25 based on our experience with the proposed model. Additionally, we augment the 3D data annotations through rotation and scaling to increase the number of samples per class for training purposes. The RBF kernel-based distance mapping presents two distinct properties that ensure a positive proportionality between the mapping and actual joint movements in 3D space. Firstly, the universal approximation property allows for the estimation of continuous functions to arbitrary precision, making it useful for representing the complex relationships within 3D sign language data. Secondly, the RBF kernel characterizes local receptive fields, ensuring a positive relational feature representation between the disjoint 3D joints on the skeleton.

Though the proposed JMAMs are capable in representing 3D skeletal data as a meaningful structure, it fails to model detailed relationships among groups of joints. Sign language users exclusively use these joint group relations to differentiate closely matched signs. Figure.4 shows signs that have more than 80% matched motion and spatial content. Hence, the second most identified que in SLR by humans is the relatedness between body parts. For example, head and hand fingers, two hands, face and fingers etc. To make these relationships between body parts significant, research has been directed towards part-based classifiers [55], [56], [57]. The results are encouraging when compared to whole body skeletal joint representations. In this context, we present an additional investigation to model these body part relationships using our JMAMs to validate the whole body JMAMs. The validation is necessary to understand the capabilities of whole body JMAMs in modelling joint relationships when compared to part based JMAMs. Moreover, the part-based models using deep neural networks require a large set of trainable parameters due to increased dataset size. The subsection gives the construction of part based JMAMs on the 3D skeletal sign language data.

C. THE JOINT GROUP TO GROUP MOTION DIRECTED JMAMS (JG2GMDMS)

The skeleton used in this work consists of 57 joints that are distributed non-uniformly across the spatial dimensions. This contrasts with the more evenly distributed skeleton used in action recognition, as can be observed in Figure.2. The upper row in Figure.2 shows the byparts skeleton models used in previous works for human action recognition, while the lower row shows the byparts of the proposed 3D sign language skeleton. By comparing the upper and lower rows in Figure.2, we conclude that the 3D skeleton in this work is non-uniformly distributed across the spatial dimension.

In the proposed jg2gMDMs model we divided the joint space into 5 regions as per our knowledge of motion in sign language data. Based on the findings on SLR, the communication between 3D joints of these regions mostly decides the sign label. These five parts are named as Head (4 – Joints) + Face (12 Joints) + Chest (3 Joints) + Left Hand (19 Joints) + Right Hand (19 Joints) = Total 3D skeletal Joints of 57. Except for hand joints, the remaining regions have non-uniform joint distributions. These regions are named as $\{J_h, J_f, J_c, J_l, J_r\} = J_p \forall p\ joints$. The suffixes are head, face, chest, left and right hand. Conventionally, the maps were constructed within the joint groups or across two joint groups. This process has resulted in good recognition accuracies

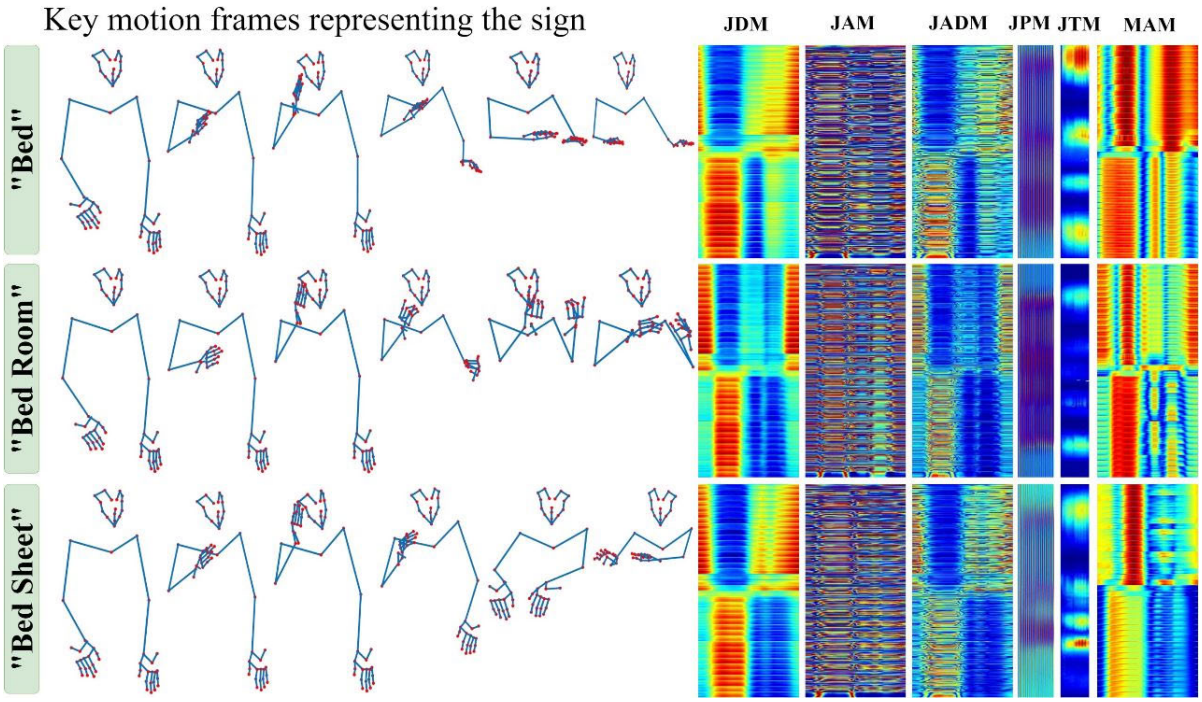

FIGURE 4. Joint movement affinity maps and comparisons with other types of feature maps.

for HAR models [21], [22]. However, in case of SLR data, the head, chest, and face are mostly stationary in all the singing process. Based on our findings, one hand is also not used more than 30% of the signs in Indian Sign Language (ISL). Hence the idea of computing JMAMs between two stationary groups such as chest and head will result in a map that is constant across all signs. More importantly, these non-varying maps does not add to the features computed during the training process for label identification.

Therefore, we propose computation of features from one group to all the remaining joint groups in the SL skeleton. In our model, the distance features are computed between a joint group and rest of the groups, or we can call all the remaining joints of the skeleton. We call our model joint group 2 group motion directed JMAMs (jg2gMDMs).

Let J_g^i be the i^{th} joint in group $g = \{[h], [f], [c], [r], [l]\}$ and $J_{g'}^j$ be the j^{th} joint in group $g' = \{f, c, r, l\}$, the Euclidian distance metric in a frame f is given as

$$\psi^{ij}(g, g') = [d_x; d_y; d_z] \epsilon R^{\left(J_g C1 \times J_{g'} C1\right) \times 3} \forall f = 1 \rightarrow F \quad (6)$$

Subsequently, the computed Euclidean distance features between the joint groups $\{\Psi_f(g, g')\}$ are converted into a group-to-group movement affinity feature matrix $M_{gg'}(\bullet)$ which can be computed using the following kernel function given as

$$M_{gg'}(\bullet) = \exp\left(-\frac{\left(\Psi_f(g, g') - \Psi_{f+1}(g, g^i)\right)}{2{\xi_{gg}}^2}\right) \forall (x, y, z, f) \quad (7)$$

Solving for Equation(7) with result in three movement affinity feature matrices in (x, y, z) directions between one group and rest of the groups prefabricated across F frames. Finally, this results in G maps given by $\{M_h, M_f, M_c, M_r, M_l\}$. The head joint group is represented by $M_h \in R^{(4C1 \times 53C1) \times 3 \times (F-1)}$. We have $4C1 \times (12C1 + 3C1 + 19C1 + 19C1)$ features in M_h as there are 4 joints in head group and 53 joints in remaining groups. The dimension 3 represents (x, y, z) axis. Similarly, we will have face, chest, right and left hand joint movement affinity feature representations as $M_f \in R^{(12C1 \times 45C1) \times 3 \times (F-1)}$, $M_c \in R^{(3C1 \times 54C1) \times 3 \times (F-1)}$, $M_r \in R^{(19C1 \times 38C1) \times 3 \times (F-1)}$ and $M_l \in R^{(19C1 \times 38C1) \times 3 \times (F-1)}$ respectively. The resulting five JET color-coded JMAMs are plotted in Figure.5 for three different signs along with their whole-body JMAMs.

The signs depicted in Figure.5 were chosen based on previous experience in [18] and [31] with JADMs trained on a deep convolutional network, where during inferencing, these signs resulted in maximum false positives and true negatives. The failure was attributed to the similarity in hand and finger movements across the three signs, which is visually evident in the whole body JMAMs in Figure.5. Similarly, training the JMAM data on the network architecture in [58] also resulted in the same problem due to the inability of CNN automated features to represent dissimilar spatiotemporal variations across signs.

Initially, this seemed like a joint-to-joint relationship representation problem, which was later addressed as a joint group to group motion directed (jg2gMDMs) approach as shown in Figure.6. The maps generated with jg2gMDMs exhibit

more variational patterns, which enable better differentiation among signs with a high level of matching content. These patterns are clearly observable in the JMAMs of groups in Figure.5 below.

D. TRAINING DATA CONSTRUCTION

Figure.1 shows the 3D joint skeleton with 57 joints used in the construction of sign language. This sign language skeletal data was developed with 3D motion capture technology with 8 motion capture camera sensors. The model was unique to our dataset and extensive research has been conducted to arrive at the skeleton shown in Figure.1. This skeleton has the capability to capture all the signs defined in Indian sign language. We used 10 different signers to construct the dataset. In each capture instance, the system captured an average of 550 frames per sign. The captured data is carefully reconstructed for any missing joints across all the signs. We named our 3D sign language dataset as developed at KL University, Biomechanics and Vision computing research centre as KL3DISL. The dataset consists of 500 daily used signs from multiple categories. To summarise, KL3DISL has 500 skeletal sign 3D labels, spanning over 25000 samples with 10 subjects repeating each sign 5 times. Additionally, the 3D skeleton can be oriented in any direction along x, y and z axis to generate cross view data for training and testing purposes. Similar data on human action recognition (KL3DHAR) is also captured with 102 actions by 10 different subjects and is available at https://www.kluniversity.in/blog/biomechanics-and-vision-computing-research-centre.html.

This 3D joint data of each sign video is transformed into one JMAM image per class. Since we have 10 subjects with 5 repetitions per class, it would be 50 JMAM images per class, of which 70% are used for training. Subsequently, training a CNN-based classifier with 35 samples per class would result in overfitting, poor generalization and difficulty in learning complex patterns in the maps even after data augmentation. Despite making a more robust feature map (JMAMs), it has become difficult to use them for generalizing a training model. In addition to the above training issues, there is a problem of JMAM image resolution. For example, a sign class having 420 video frames would result in a JMAM of size $\frac{57\times(57-1)}{2} \times (420-1) \times 3$ i.e., $1596\times419\times3$. This means that the width-to-height ratio in JMAMs is noticeable. In order to apply JMAM images with the above resolution requires the precise design of layers in a neural network. Hence, we scaled down the JMAM image with unstructured resolution to structured one that is acceptable to standard networks such as VGG-16 and RESNET50. However, when scaling and normalization were applied to the original JMAM images for training standard deep network models, they failed to generalize no more than 34% of the class labels.

The above challenge was addressed by developing a new multi-stream multi-resolution convolutional neural network architecture with a spatial attention model for both JMAMs and jg2gMDMs. In order to address the training issues in JMAMs and jg2gMDMs, a detailed architecture of the proposed multi-resolution convolutional neural network with spatial attention (MRCNNSA) is provided in the following section.

E. MULTI-RESOLUTION CONVOLUTIONAL NEURAL NETWORK WITH SPATIAL ATTENTION (MRCNNSA)

This section describes a new deep network architecture that is suitable for handling JMAMs dataset with unstructured resolution. The goal of this network design is to classify 3D sign language based on the features of JMAMs at multiple structured resolutions in place of single unstructured resolution. Contemplating on the past experiences on the derived feature dataset, we present multi resolution CNN with spatial attention (MRCNNSA) as shown in Figure.6. This network has been designed to overcome the problem of overfitting and generalization in the previous models due to scaling and data shortages. The MRCNNSA is multi stream CNN backbone architecture with distributed attention across different resolutions of the input JMAMs. The lower resolution features extracted are more likely to represent global patterns and low frequency textures in the image data. The higher resolution features represent high frequency contours and fine-grained textures. Since feature data (JMAMs) is of size $1596\times419\times3$, it is difficult to standardize a fixed scale image for learning all the patterns in the image. Specifically, the JMAM feature map representing 3D joint spatial and temporal distributions of a skeletal sign is unstructured across its dimensions. Resizing JMAM to standard resolutions (224×224) to suite the input dimensions of deep network architectures like VGG or ResNet which subsequently learns only global patterns and low frequency textures. This resulted in missing the high frequency and fine-grained texture patterns representing finger joint movements. For higher resolution JMAMs the low frequency and global features representing the hand and head joint movements were compromised. Simultaneously learning both the low frequency hand and high frequency finger joint features in JMAMs is a tough task for standard networks that train on fixed input image resolution. Hence, we propose MRCNNSA, which is designed with 5 streams of convolutional layers accepting input at multiple resolutions. They are $64\times64\times3$, $128\times128\times3$, $256\times256\times3$, $512\times512\times3$, and $1024\times1024\times3$. Bi-Cubic interpolation method was applied on the original $1596\times419\times3$ to convert JMAMs of multiple resolutions.

The layer composition in each stream is the same except for the input receptive field, which changes based on the input across each stream. All convolution filters in all layers and streams are of size 3×3 with stride 1 and border retention. All layers have the activation function ReLu. Some convolutional layers are followed by 2D maximum pooling with stride 2, transforming the resolution to half. The primary objective of this network is to address the non-uniform resolution in JMAMs, which has been a difficult proposition during the training process. The MRCNNSA in Figure.6

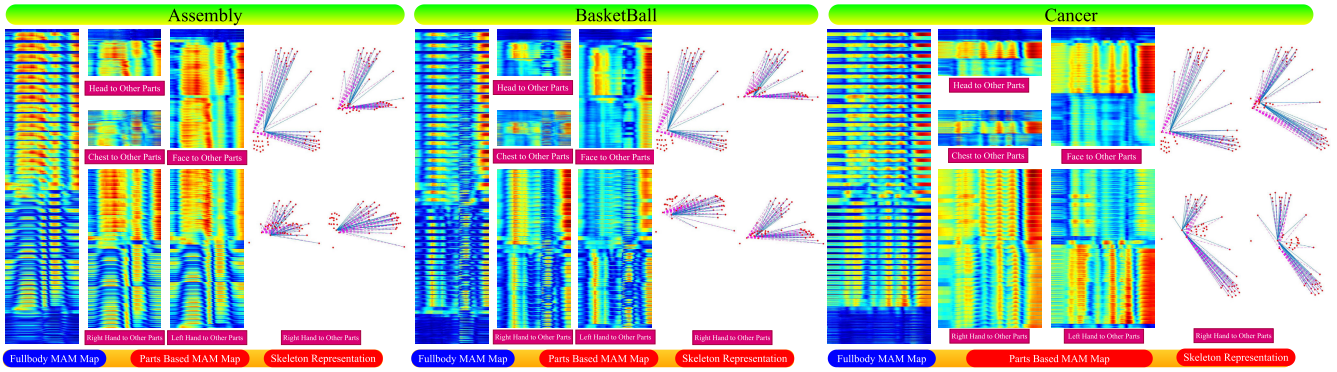

FIGURE 5. The byparts JMAMs along with full body JMAM for signs from ISL on 3D data.

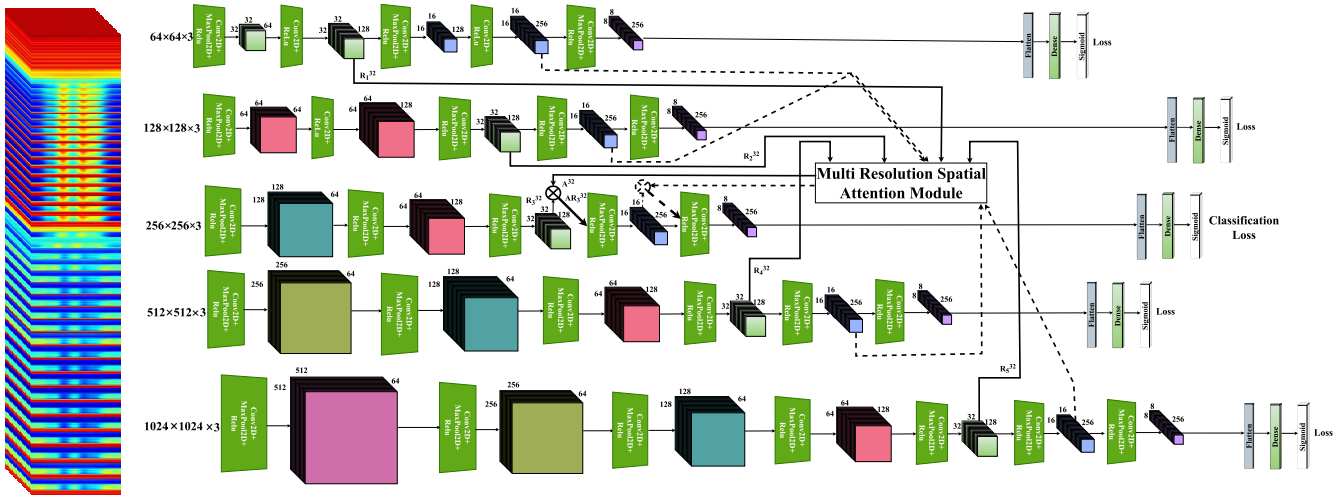

FIGURE 6. Architecture of multi-resolution convolutional neural network with spatial attention (MRCNNSA) for JMAMs.

shows the low-resolution features (64,128)$\left[R_1^{32}, R_2^{32}\right]$ and high-resolution features (512,1024) $\left[R_4^{32}, R_5^{32}\right]$are combined into mid-range resolution of 256 $\left[R_3^{32}\right]$. The combination of $\left[R_1^{32}, R_2^{32}, R_4^{32}, R_5^{32}\right]$ and R_3^{32} in Multi Resolution Spatial Attention Module (MRSAM) will result in A_3^{32}. This will be applied to mainstream network as AR_3^{32} as

$$AR_3^{32} = \prod_{\forall features} A_3^{32} R_3^{32} \tag{8}$$

The resulting AR_3^{32} combines the high-resolution features such as high-level semantic information with the low-resolution low-level texture information using the MRSAM. The combination can be across any level in the middle CNN stream at locations defined by $AR_3^{64}, AR_3^{16}, AR_3^{8}$. The MRSAM is shown in Figure.7 along with the traditional channel and spatial attention modules. The objective of modifying spatial attention in Figure.7(b) into Figure.7(d) is to reduce dimensionality of features in the intermediate layers. The output of attention model if Figure.7(d) can be formulated as

$$\begin{aligned} \tilde{X}_{32}^{R3} &= X_{32}^{R3} \otimes X_1^{R1245} \\ X_1^{R1245} &= f_w\left(f_w\left(X_2^{R12}\right)\right) \otimes f_w\left(f_w\left(X_2^{R45}\right)\right) \\ &\otimes f_w\left(f_w\left(X_2^{R24}\right)\right) \end{aligned} \tag{9}$$

Here $f_w(\bullet)$ denotes the features extracted by the 1×1 convolution operator which reduces the dimensionality of the features in the output of the attention network. The w is the weight operator generated during the training of MRSAM. The terms $X_{32}^{R1}, X_{32}^{R2}, X_{32}^{R3}, X_{32}^{R4}, X_{32}^{R5}$ represent 32 features. The term X_{32}^{R3} represents features of R3, which is the middle stream in Figure.6. Similarly, terms R1, R2, R4 and R5 represent streams in Figure.6. The outputs after the second set of 1×1 convolutions are

$$\begin{aligned} X_4^{R12} &= \left(f_w\left(X_{32}^{R1}\right) \otimes f_w\left(X_{32}^{R2}\right)\right) \\ X_4^{R24} &= \left(f_w\left(X_{32}^{R2}\right) \otimes f_w\left(X_{32}^{R4}\right)\right) \\ X_4^{R45} &= \left(f_w\left(X_{32}^{R4}\right) \otimes f_w\left(X_{32}^{R5}\right)\right) \end{aligned} \tag{10}$$

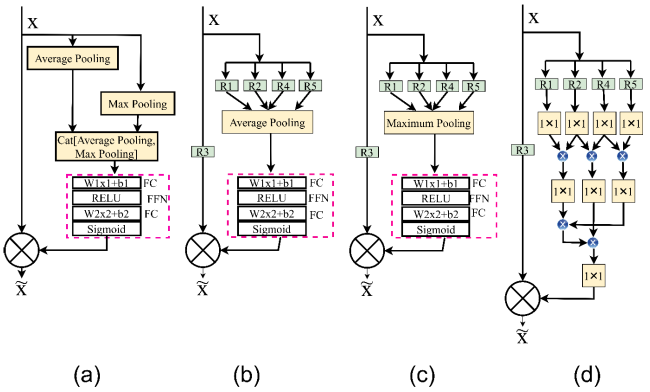

FIGURE 7. Spatial attention models. (a) Traditional spatial and channel attention network, (b) Multi resolution spatial attention network, (c) Multi – resolution channel attention network, and (d) Proposed multi – resolution spatial attention network (MRSAM).

where X_4^{R12} are the dimensionally reduced features after 1×1 convolution layer and the piecewise multiplier $\otimes$. Similarly, X_1^{R1245} gives the attention map generated from the multi resolution features of dimension 32. Finally, $\tilde{X}_{32}^{R3}$ denote the attention features that are propagated back into the R3 stream in Figure.6 with dimension 32. Additionally, the 64-, 16- and 8-dimension features in stream R3 of Figure.6 can also be influenced using the MRSAM network. This can be done either independently or all at once. We will be showing the results of this testing in the experimentation section.

Accordingly, the convolution operations in the network in Figure.6 are formulates as

$$\Theta_{R_\alpha} = \arg\min_{\Theta_{R_\alpha}} L_\alpha \left(\Theta_{R_\alpha} : I_\alpha (x), y\right) \tag{11}$$

where $I(x)$ gives the JMAMs and the model $m\left(\Theta_{R_\alpha}\right)$ where Θ is a set of trainable parameters in each of the resolution R streams denoted by $\alpha \forall 1$ *to* 5. The aim of each stream in Figure.6 is to extract spatial features from JMAMs in multiple resolutions as shown in Figure.6. However, the R3 stream in Figure.8 integrates the global features from $I_{256}(x)$ with attention features extracted from $I_{64}(x), I_{128}(x), I_{512}(x), I_{1024}(x)$ JMAMs. Each stream is trained separately with trainable parameters $\Theta_{R_\alpha} \forall \alpha = 1$ *to* 5 by optimizing the loss function L_α on the entire dataset. The L_1, L_2, L_3, L_4 are the local independent losses and L_3 is the categorical cross entropy loss to classify labels represented by y.

The trained models $m\left(\Theta_{R_\alpha}\right) \forall \alpha = 1, 2, 3, 4, 5$ outputs a set of spatial features $\left\{X_\alpha^l\right\} \forall l \rightarrow$ *layer index in stream* α representing the JMAMs at l^{th} layer in α^{th} stream as

$$X_\alpha^l = \sum_{i=1}^{C} \sum_{j=1}^{C} I_n (i, j) K ((k - i)(k - j)) \forall l, \alpha, n \tag{12}$$

where K is the kernel size across each of the layers. The$N \times k_{dense}$ features input the dense layer in R3 for classification. The convolutional layers in all streams use rectified linear unit defined as

$$R(z) = \max(0, z) \tag{13}$$

where z is the output of the neuron. Similarly, the dense layers have tanh and the SoftMax layer has sigmoid activation defined as

$$\sigma(z) = \frac{1}{\left(1 + e^{-z}\right)} \tag{14}$$

The above representation is applied further in the training of similar network build for the classification of KL3DISL with jg2gMDMs.

F. JG2GMDMS CLASSIFIER NETWORK

Specifically, to build a classifier network similar to the above MRCNNSA on jg2gMDMs, we have to decrease the input image resolutions as the maps are of smaller dimension when compared to JMAMs. The dataset invokes a 3- stream network shown in Figure.8 for each body part feature map. Specifically, the MRCNNSA for body part-based system has 15 streams that are clustered into 5 bundles $\{b_h, b_f, b_c, b_l, b_r\}$ namely, "Head_to_Other_parts", "Face_to_Other_Parts", "Chest_to_Other_Parts", "Left_Hand_to_Other_Parts" and "Right_Hand_to_Other_Parts" given as $\{M_h, M_f, M_c, M_r, M_l\}$. A bundle b_f accepts one body part relational color-coded map M_f as input in three different resolutions. Since the body part relational maps are smaller, the resolutions selected are 16, 32, and 64. In b_f the R_1 and R_3 streams are trained for features in the 1^{st} training instance. The 2^{nd} instance trains the $b_f's$ R_2 stream by fusing multi resolution attention feature A^8 with inference features $\left[R_1^8, R_3^8\right]$ as AR^8. Though the R_3 stream is trained as a classifier, it is used as a feature extractor of head features $f_h \in R^{1024 \times 1}$. Similarly, for all other body part, we have $\{f_f, f_c, f_l, f_r\} \in R^{1024 \times 1}$ extracted from the output of last convolution layer. Finally, the 3^{rd} training instance trains a fully connected dense layers to estimate the pose from concatenated all body part features $f_c = \{f_h, f_f, f_c, f_l, f_r\} \in R^{5120 \times 1}$. Consequently, each stream in the network of Figure.8 has to be trained independently for separate body parts and the resulting features are concatenated for classification. The concatenated body part features are trained using the model in Figure.9.

The latency in the detection process due to the establishment of JMAMs for the captured 3D pose and translating that into the motion capture 3D JMAMs is major limitation for real-time SLR. The trained system takes approximately 6 to 8 seconds to display the recognized sign text on the inferencing video, which is a significant challenge for real-time implementation. To reduce this latency to less than 2 seconds, dedicated hardware and intensive training of the models are required. The penultimate section of the paper presents experiments conducted on the above datasets with the proposed methodology to evaluate the performance of the model.

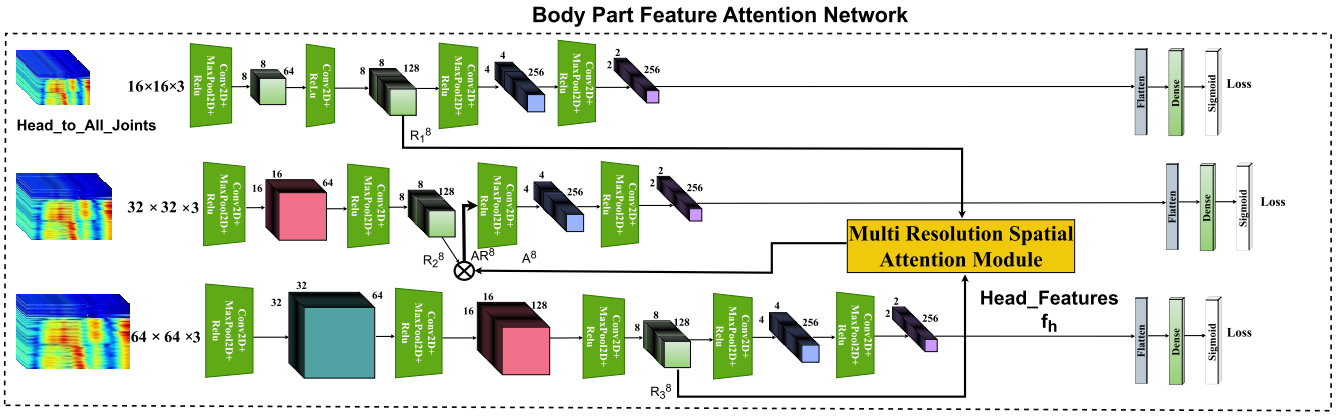

FIGURE 8. Architecture of body part-based multi-resolution convolutional neural network with spatial attention (MRCNNSA) for jg2gMDMs.

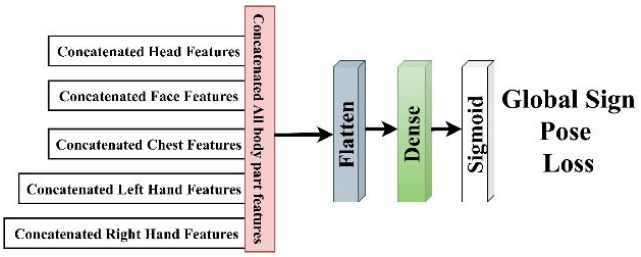

FIGURE 9. Last layers in MRCNNSA for Jg2gMDM inputs for body part-based 3D ISL recognition.

IV. EXPERIMENTATION AND DISCUSSION

This section details the implementation of the proposed models on the datasets and metrics used for evaluation. Additionally, comparisons are drawn against multiple standard CNN architectures as backbone networks using benchmark datasets and state-of-the-art sign language (action) recognition models. Furthermore, ablation analysis for the attention model at multiple feature resolutions in the classifier is presented.

A. KL3DISL AND BENCHMARK DATASETS

The proposed methods were specifically designed for 3D skeletal datasets, with KL3DISL [44], [59], being extensively used throughout the discussion and as input for classification tasks on the proposed MRCNNSA. To test the generalization capabilities of the two networks proposed in this work, we selected four challenging 3D skeletal datasets: NTU RGBD [28], KL3DYOGA [23], CMU [29] & HDM05 [30]. As no 3D sign language data is available from either a 3D motion capture or a Kinect system, we chose 3D human action datasets instead. NTURGBD contains 120 classes in 114,480 3D skeleton videos of human actions captured using a Kinect sensor. KL3DYOGA is our own dataset of 42 yoga poses in 16,800 3D skeletal videos. CMU and HDM05 have 23 and 70 action classes distributed across 2,605 and 700 3D skeletal videos respectively.

B. EVALUATION METRICS

ITo evaluate the performance of the proposed models on 3D sign language recognition using color-coded spatiotemporal features, we employ the following metrics: overall accuracy (OA), class-specific accuracy (CSA), Cohen's kappa coefficient, and top–1% accuracy. These metrics are formulated as: (15), shown at the bottom of the next page,

$$\text{CSA}_\text{i} = \frac{\text{True Positives for } i^{th} \text{ class}}{\text{Total } i^{th} \text{ class test samples}} \tag{16}$$

$$\text{Cohen}'\text{s Kappa Coefficient k} = \frac{P_o - P_c}{1 - P_c} \tag{17}$$

where

$$P_o = \frac{\text{Sum of diagonal elements in Confusion matrix}}{\text{Total number of instances}} \tag{18}$$

$$P_c = \frac{\text{sum of row totals} \times \text{sum of column totals}}{(\text{Total Number of Instances})^2} \tag{19}$$

Overall accuracy (OA) defined as the ratio of instances classified correctly against all the dataset instances gives the complete performance of the model on the test samples.

C. IMPLEMENTATION OF MRCNNSA FOR WHOLE AND GROUPED BODY JMAMS

This subsection presents the implementation process of MRCNNSA for JMAMs and jg2gMDMs on the 5 skeletal datasets considered in this work. The number of joints in each dataset could arise dimensionality problems, with KL3DISL having 57 joints across 400 frames for most signs and KL3DYOGA being a full body skeleton of 39 joints. The benchmark datasets NTURGBD, CMU and HDM05 have 25, 41, and 31 joints, respectively. Another challenge is the number of frames per video sequence, with KL3DISL having a range of 330 to 450 frames, KL3DYOGA having around 4200 to 5600 frames per sample yoga pose and the rest falling in the normal category with 430 to 670 frames. The JMAMs distance metric between joint pairs and color coding of 3D data across datasets KL3DISL, KL3DYOGA, NTURGBD,

CMU and HDM05 has resulted in a JMAM of size $1596\times419\times3$, $741\times4331\times3$, $300\times521\times3$, $820\times432\times3$ and $465\times523\times3$ respectively. Applying the JMAMs with their original resolutions was difficult due to the standardization of deep learning architectures and specifically it is challenging to control the layer dimensionalities. Hence, the traditional approach followed was to resize to the shape acceptable to the network. This size is $224\times224\times3$ or $256\times256\times3$. The proven networks are VGG-16, ResNet50, InceptionV4, and MobileNet. When whole-body JMAMs of KL3DISL were trained and tested on these standard networks, the top 1% accuracy (T1%A) was 43.85% and a top 1% error rate of 0.3521 when whole-body JMAMs of KL3DISL were trained and tested on them.

As discussed in the literature, attention has been shown to enhance the overall classification accuracies on various types of datasets. In this study, we utilize the standard high-performing attention models listed in Table.2 to train and evaluate the JMAMs on the considered datasets.

The results in Table.2 demonstrate an improvement over that of Table.1, indicating that the attention module or network is effective for analysing the feature rich images representing spatiotemporal 3D data of KL3DISL and other datasets. However, the top 1% error rate remains high and when analyzed on individual classes, the average error rate was above 40% for KL3DISL. Further analysis revealed that the above networks and their intermediate layered feature maps struggle to learn variations that occur in a short frame count, such as a sign or action that quickly changes over time and is captured by very few frames. For instance, in the signs 'Bed Sheet' and 'Bedroom' shown in Figure.4, the pattern in pixel variations is almost identical, causing most of the above networks to misclassify KL3DISL. After conducting multiple experiments and analyses on attention-based networks, we developed the MRCNNSA model in Figure.6 for JMAMs and the jg2gMDMs models in Figure.8. We adopted a multi-resolution input in multiple streams of CNNs to capture both local pixel changes and preserve high-frequency semantic features until the dense layers.

It takes multi-resolution input across independent multiple streams for capturing both low- and high-level features that are the basis for generating a spatial attention map. Subsequently, this spatial attention map is fused at matching dimensions in the mid-resolution input stream to generate wholistic features across a sign (action) label. Implementation details of the networks in Figures 6, and 8 are given in Table.3.

The comparison in Table.3 demonstrates that our proposed models have fewer trainable parameters than standard architectures like VGG16 (138M), ResNet50 (23M) and InceptionV4 (43M). The allowable inferencing loss is the prediction loss that can be tolerated during feature extraction from the trained models. The training process is implemented in two phases. In the first phase, models in streams 1, 2, 4, and 5 are trained using the parameters in Table.3 on input JMAMs. These trained models are then saved on disk. In the second phase, the model in stream 3 is trained along with the attention network and the saved trained networks in streams 1, 2, 4, and 5. These trained networks are then inferenced to extract features required for generating the attention maps for the backbone network. At this stage, the network becomes end-to-end trainable.

In contrast, the total number of parameters in the second phase of training is higher than standard networks. However, this does not increase the computation time or memory usage in our MRCNNSA model for both JMAMs and jg2gMDMs, as 80% of the parameters are already trained and only extract features at multiple resolutions. Although the total number of parameters in MRCNNSA is 92M, out of which only 19M are trained. These 19M parameters are in the 256-resolution middle stream along with the layers in the attention network. Similarly, the networks in Figure.8 and Figure.9 for jg2gMDMs are trained separately using the same approach. Finally, we apply the same procedure to all other benchmark datasets used in this work.

The trained models on JMAMs and jg2gMDMs are evaluated on KL3DISL, KL3DYOGA, NTURGBD, CMU, and HDM05 in the following section.

D. JMAMS VS OTHER COLOR CODED FEATURES

The objective of this section is to identify whether the hypothesis given in the introduction about the JMAMs is true. Specifically, this section provides the comparative analytics of testing JMAMs on trained MRCNNSA for our KL3DISL and other benchmark datasets. 20% of test data in all datasets is used to evaluate the MRCNNSA. The metrics used are OA, CSA, and Top 1% accuracy. The dimension of spatial attention features in the trained model MRCNNSA is $32\times32\times128$. This dimension gets reduced due to averaging in the attention network to $32\times32\times1$, which is then multiplied to the 3^{rd} stream or main classification stream as shown in Figure.8. Undoubtedly the training and the classification accuracy were better than early or late fusion. The results of this experiment were organized in Table.4. Obviously, the scores on NTURGBD should be reasonably lower than all other datasets as it is a markerless system. However, we found HDM05 at the lower end of performance due to many missing joints when compared to NTURGBD. The percentage of missing joints in KL3DISL and KL3DYOGA is less than 0.1% across all samples. It is customary to reconstruct the missing joints from two adjacent joints by averaging their position vectors. Except for the number of output labels for

$$OA = \frac{\text{Number of correct Predictions across all classes in the dataset}}{\text{Total number of predictions}} \tag{15}$$

TABLE 1. JMAMs and jg2gMDMs on standard networks.

Standard Top Rated CNN Networks for Image Classification	Top 1% Accuracy JMAMs (Whole Body)			Top 1% error rates			Top – 1% Accuracy jg2gMDMs			Top 1% error rates		
Datasets	KL3DISL	CMU	NTURGBD	KL3DISL	CMU	NTURGBD	KL3DISL	CMU	NTURGBD	KL3DISL	CMU	NTURGBD
VGG-16	37.58%	39.45%	38.45%	0.4518	0.4204	0.4427	38.19%	41.14%	39.11%	0.4389	0.4161	0.4225
VGG-19	39.97%	43.56%	41.73%	0.4489	0.4095	0.4188	40.69%	45.83%	42.32%	0.4125	0.3948	0.4072
ResNet34	48.59%	50.89%	50.14%	0.4176	0.3845	0.3971	50.44%	52.05%	51.32%	0.3936	0.3735	0.3896
ResNet50	52.12%	56.46%	54.84%	0.3921	0.3521	0.3698	56.52%	59.09%	57.20%	0.3493	0.3147	0.3215
ResNet101	50.24%	55.12%	51.99%	0.4097	0.3772	0.3809	53.72%	56.80%	54.59%	0.3849	0.3509	0.3618
Inception V3	41.67%	44.52%	43.07%	0.4212	0.4042	0.4097	42.96%	46.39%	44.09%	0.3974	0.3695	0.3714
InceptionV4	43.11%	45.09%	44.46%	0.4111	0.3918	0.4015	43.98%	46.73%	45.38%	0.3859	0.3509	0.3601
MobileNetV3Lite	44.45%	47.13%	46.01%	0.4032	0.3875	0.3946	46.02%	49.41%	47.92%	0.3703	0.3413	0.3447
MobileNetV3QAT	43.97%	46.47%	44.83%	0.4101	0.3711	0.4075	44.47%	48.55%	46.45%	0.3882	0.3552	0.3663

TABLE 2. Top image classifier architectures with attention module from past works that were trained from scratch on the data in this work.

Standard Top Rated CNN Networks for Image Classification	Top 1% Accuracy JMAMs (Whole Body)			Top 1% error rates			Top – 1% Accuracy jg2gMDMs			Top 1% error rates		
Datasets	KL3DISL	CMU	NTURGBD	KL3DISL	CMU	NTURGBD	KL3DISL	CMU	NTURGBD	KL3DISL	CMU	NTURGBD
ViT – Vision Transformers	62.97	65.99	64.75	0.3967	0.3652	0.3752	66.98	68.04	67.45	0.3778	0.3548	0.3662
DeiT – Data-efficient Image Transformers	63.91	66.78	65.48	0.3892	0.3583	0.3618	67.27	69.57	68.93	0.3695	0.3368	0.3473
SENet – Squeeze-and-Excitation Networks	59.4	61.76	62.19	0.3905	0.3722	0.3814	62.52	65.02	64.79	0.3743	0.3354	0.3424
CBAM – Convolutional Block Attention Module	56.78	58.15	58.26	0.3895	0.3525	0.3871	59.76	60.58	60.07	0.3687	0.3409	0.3479
BAM – Bottleneck Attention Module	58.25	60.65	58.82	0.3834	0.3681	0.3748	61.29	63.85	61.92	0.3554	0.3208	0.3274
ECA-Net – Efficient Channel Attention Networks	56.04	57.06	56.78	0.4014	0.3839	0.3892	58.98	60.07	59.15	0.3815	0.3512	0.3528
CoAtNet – Convolutional Attention Networks	67.47	69.4	68.13	0.3724	0.3344	0.3513	70.33	71.88	70.97	0.3318	0.3021	0.3086

TABLE 3. Implementation details of the networks in Figure's 6 and 8.

Model	Input Type	Input Size	Learning Rate	Trainable Parameters	Hyperparameter Initialization	Allowable Inferencing Loss
MRCNNSA (Figure.6)	JMAMs	64	0.01	18.1M	Gaussian Distribution with mean 0 and variance 1.	0.01
		128	0.01	18.23M		0.01
		256	0.001	18.39M		0.0001
		512	0.001	18.43M		0.001
		1024	0.0001	18.61M		0.001
MRCNNSA (Figure.8)	jg2gMDMs	16	0.1	1.74M		0.1
		32	0.1	1.75M		0.001
		64	0.01	1.89M		0.01
Multi–Resolution Spatial attention network (MRSAM).	JMAM Features	32	Same as the learning rate of the Backbone network using JMAM input	1K (4)	-	-
				0.5K(3)		
				0.1K(1)		
	Jg2gMDM Features	8	Same as the learning rate of the Backbone network using jg2gMDM input.	1K (4)	-	-
				0.5K(3)		
				0.1K(1)		

each of the considered datasets, all other hyperparameters were kept constant.

However, there are cases in our KL3DISL dataset where accuracy is very high for simple signs and very low for complex signs shown in Figure.4. Hence, class specific accuracy (CSA) has been selected as a metric to evaluate the proposed classifier. It is the accuracy computed for instances within a class giving the label wise performance of the model. In Table.4, the CSA has been computed for individual classes in the test dataset and was presented as an average value. The last metric is Top 1% accuracy which measures the best OA achieved by the model during inferencing on a single sample data other than the dataset.

At this stage, it would be interesting to interpret why JMAMs performed better when compared to all other maps. The reason for this has been interpreted based on the Equation(6) where there is an inverse relation between joint distances ψ and movement affinity features M. Consequently, this type of expression has been shown to provide excellent dynamic range in colour coded movement affinity matrix M. However, true this might look it would be interesting to explain why the MRCNNSA recorded better OAs than the standard architectures in Table.1. Subsequently, the attention mechanisms in Table.2 that were proposed earlier are enough for KL3DISL JMAMs without the model in Figure.9. To explain MRCNNSA learning model performance, we propose to apply Cohen's kappa coefficient (κ). It measures level of agreement between two classifiers in a classification task. Here we perform inferencing on test dataset consisting of 5 samples per class on only KL3DISL dataset. The ResNet50 model in Table.1 are used with our proposed attention mechanism as has the highest accuracy over other models. The ResNet50 model architecture in Table.1 follows that of Figure.8. The middle stream gets an attention map for fusion into the main feature stream at the 3^{rd} RES block with a resolution of 32. Table.5 gives the (κ) between our MRCNNSA and the ResNet50SA in Table.1 as backbone.

The results in Table.5 significantly show the quality of the JMAMs used for the 3D skeletal classification tasks. In the meantime, (κ) for other maps on the above networks were also computed to ascertain their impact on the recognition of 3D sign language. The JDM ($\kappa = 0.423$), JADM ($\kappa = 0.526$), JTM ($\kappa = 0.481$) and JVM($\kappa = 0.532$). The values show that the models are in moderation in inferencing signs using these feature maps. However, the JMAMs recorded ($\kappa = 0.763$) between MRCNNSA and ResNet50SA, which means there is a substantial agreement between the two in discovering 3D signs. This metric shows

TABLE 4. Performance of the proposed JMAMs against other types of color-coded features.

Datasets	KL3DISL			KL3DYOGA			CMU			HDM05			NTURGBD		
Metrics/Maps	OA	CSA	Top 1%	OA	CSA	Top 1%	OA	CSA	Top 1%	OA	CSA	Top 1%	OA	CSA	Top 1%
JDM	71.47	75.04	53.92	72.54	76.89	54.73	75.04	79.24	56.91	72.78	77.18	55.03	72.41	76.74	54.69
JAM	74.18	78.48	55.96	75.29	80.18	56.81	77.88	82.22	59.06	75.54	80.41	57.11	75.15	80.04	56.76
JADM	77.25	81.34	59.79	78.41	83.03	60.69	81.11	85.49	63.1	78.67	83.33	61.02	78.27	82.89	60.64
JPM	75.97	79.76	59.55	77.11	82.15	60.44	79.76	84.04	62.85	77.36	82.14	60.77	76.96	81.91	60.4
JTM	67.54	71.59	50.96	68.55	73.27	51.73	70.91	75.79	53.78	68.78	73.52	52.01	68.42	73.49	51.66
JVM	75.69	80.31	57.84	76.82	81.58	58.71	79.47	84.14	61.05	77.08	81.86	59.03	76.68	81.31	58.67
QJVM	79.46	83.43	60.73	80.65	86.05	61.63	83.43	89.81	64.09	80.92	86.36	61.97	80.51	85.01	61.59
JMAM	87.34	92.58	64.19	88.65	94.84	65.16	90.86	97.14	67.11	88.13	94.27	64.94	87.67	88.38	64.51

TABLE 5. The Cohen's kappa coefficient (κ) for comparing the similarities between the proposed MRCNNSA and the top performing state-of-the-art ResNet50.

ResNet50SA/ MRCNNSA	Bed	Bed Sheet	Bedroom	Bathroom	Drawing room	Iron sheets	Paper Sheets	Head	Hair	Eyes	Eye Brows	Woamn	Mother	Daughter
Bed	130	11	5	4	0	0	0	0	0	0	0	0	0	0
Bed Sheet	2	143	2	2	0	0	0	1	0	0	0	0	0	0
Bedroom	4	5	139	2	0	0	0	0	0	0	0	0	0	0
Bathroom	2	8	4	136	0	0	0	0	0	0	0	0	0	0
Drawing Room	0	0	3	0	143	0	2	0	1	0	1	0	0	0
Iron Sheets	0	0	0	0	0	141	7	1	0	0	0	0	0	1
Paper sheets	0	0	0	0	2	11	136	1	0	0	0	0	0	0
Head	0	0	1	0	0	0	0	135	13	0	1	0	0	0
Hair	0	0	0	0	0	1	0	17	132	0	0	0	0	0
Eyes	0	0	0	0	0	0	0	3	0	127	17	1	2	0
Eyebrows	0	0	0	0	0	0	0	1	0	14	131	1	3	0
Woman	0	0	0	0	0	0	0	0	0	3	1	138	2	6
Mother	0	0	0	0	0	0	0	0	0	2	0	7	137	4
Daughter	0	0	0	0	0	0	0	1	0	1	1	4	12	131

that there is a substantial impact of JMAMs, and the multi resolution spatial attention module attached to the deep network, MRCNNSA.

On the contrary, sign language is specifically decoded using the relationships between the fingers in hands, hands with each other, and hands with head and torso. Now the JMAMs proposed have significantly improved the OA by 7.12% over the other maps in practice. However, are the JMAMs truly exploiting these relationships at a global scale in a better manner than the others. That is why the joint group-to-group motion-directed maps (jg2gMDMs) were constructed.

E. JG2GMDMS VS JMAMS

It is certain and has been proven that 3D joint contextual information will enhance the recognition accuracy of skeletal-based action recognition systems [19], [56]. However, can the JMAMs compete with the body part relational models in characterizing this information on our proposed MRCNNSA This challenge is formulated as a test accuracy comparison problem. The experiment starts by training the MRCNNSA in Figure.8 with the body part maps Jg2gMDM. The training and testing procedure is described in section III-F. The model evaluation results are shown in Table.6 for all the datasets used in this work along with the maps.

The overall accuracy is computed on the test dataset. On the whole the results in Table.6 show that the body part based relational features characterize the 3D joint information more effectively when compared to whole body maps. The difference in OA between the two models is significantly large for all maps except JMAMs. Explicitly this gap in OA is found to be around 21%. However, the OA gap for JMAMs and jg2gMDMs is lower than 6%. This meant that JMAMs have rich feature representation of 3D joint motions which are close to the relational features jg2gMDMs. Interestingly, JMAMs take few training instances when compared to jg2gMDMs. However, the Jg2gMDM based learning model uses less computational power when compared to JMAMs. This fact can be verified by the trainable parameter's column in Table.3. The next subsection validates the proposed method against the state-of-the-arts in learning systems for 3D skeletal sign(action) recognition.

F. OUR PROPOSED METHOD (JMAMS+MRCNNSA) VS STATE-OF-THE-ARTS

It was indeed difficult to benchmark the proposed method with publicly available 3D skeletal SLR datasets. However, availability of similar datasets in the domain of 3D skeletal action recognition has indeed helped the validation of the proposed model. The objective is to validate features and its corresponding colour coded maps simultaneously. In Table.7, we present the features used by some state-of-the-art methods on sign(action) recognition methods. Here 3D skeletal sign language methods are incepted from our previous works [18], [31], [44]. The comparisons were drawn using OA, the OA projected in Table.7 is obtained by the following the methods in those works on the dataset considered. Subsequently, comparison was drawn on the methods developed using colour coded feature maps and deep learning methods in Table.7. Though they used other datasets in their works, we were interested in the results obtained using NTURGBD, CMU and HDM05 for action recognition. Accordingly, we also used our previous methods on 3D skeletal sign language with 200 classes for evaluating the proposed methods. The results in Table.7 show that the

TABLE 6. JMAMs vs Jg2gMAMs: Result analysis on all the joint feature maps on the proposed MRCNNSA models for whole and part body relational joints in Figure.6 and Figure.8.

Datasets	KL3DISL			KL3DYOGA			CMU			HDM05			NTURGBD		
Metrics/Maps	OA Whole	OA Part Body	Top 1%	OA Whole	OA Part Body	Top 1%	OA Whole	OA Part Body	Top 1%	OA Whole	OA Part Body	Top 1%	OA Whole	OA Part Body	Top 1%
JDM	71.47	80.76	55.03	72.54	81.97	55.85	75.04	84.04	57.78	72.78	81.51	56.04	72.41	81.09	55.75
JAM	74.18	83.08	57.11	75.29	84.32	57.96	77.88	87.22	59.96	75.54	84.61	58.16	75.15	84.16	57.86
JADM	77.25	88.83	61.02	78.41	90.16	61.93	81.11	90.03	64.07	78.67	87.32	62.14	78.27	86.87	61.82
JPM	75.97	85.08	60.77	77.11	86.35	61.68	79.76	88.53	63.81	77.36	85.87	61.89	76.96	85.43	61.57
JTM	67.54	75.64	52.01	68.55	76.77	52.79	70.91	80.12	54.6	68.78	77.72	52.96	68.42	77.31	52.67
JVM	75.69	84.01	59.03	76.82	85.27	59.91	79.47	88.69	61.98	77.08	86.02	60.12	76.68	85.58	59.81
QJVM	79.46	88.26	61.97	80.65	89.58	62.89	83.43	92.61	65.07	80.92	89.83	63.11	80.51	89.36	62.79
JMAM	87.34	91.71	65.51	88.65	93.08	66.49	90.86	95.43	68.14	88.13	92.56	66.09	87.67	92.08	65.75

TABLE 7. Skeleton based action recognition on various datasets with their respective features.

Method	Features	Classifier	NTU-RGB + D	HDM05	CMU	KL3DISL
Skeleton based Features						
W. Chan., 2020 [60]	Joint & Bone	GCN	90.4	-	-	-
J. Dong., 2020 [49]	Joints, Bone, velocity of Joints and Bone, distance, Acceleration of Joints and Bone	GCN	90.5	-	-	-
J. Zhu., 2019 [61]	Node pairs and edge pairs	Convolutional relation network	86.2	-	-	-
W. Ding., 2020 [62]	Relations among the disjoint and distant joints	GCN	85.8	-	-	-
H. Wang., 2018 [52]	Joints, Edges, Surfaces	BiLSTM	79.5	-	86.1	-
X. Chen., 2015 [63]	Normalized 3D joint positions	ELM	-	96.7	-	-
L. Wang., 2018 [64]	Joints	LSTM	80.9	-	-	-
Y. Xu., 2018 [65]	Joints	1D CNN	85.1	-	-	-
S. Zhang., 2018 [46]	10 Different geometric features	LSTM	76.43	-	-	-
J. Liu., 2018 [66]	Tree structure joints	LSTM	69.2	-	-	-
Skeleton feature Maps						
C. Li., 2017 [21]	Joint Distance Maps	CNN	76.2	-	-	-
E. K. Kumar., 2018 [44]	Joint Angular maps	CF-ResNet CNN	75.63	81.93	76.23	-
P. Wang., 2018 [47]	Joint Trajectory Maps	CNN	76.32	-	-	-
L. Jian., 2019 [67]	Joint Location, Velocity Maps	Inception-ResNet CNN	81.3	-	-	-
T. K. K. Maddala., 2019 [23]	Joint Angular Displacement Maps	CNN	-	89.95	89.56	-
J. Ren., 2018 [68]	Displacement of Joints, Angles & Distance Maps	CNN	76.1	-	-	-
Bo Li., 2017 [69]	Translation-Scale Invariant Maps	CNN	85.02	-	-	-
S. Laraba., 2019 [70]	Joint coordinates Maps	CNN	82.07	-	-	-
K. V. Prasad., 2019 [71]	Transformed Joint Location Maps	CNN	82.37	-	-	-
C. Caetano., 2019 [72]	Tree Structure Reference Joints Maps	CNN	73.3	-	-	-
M. Liu., 2017 [73]	Motion Enhancement Maps	CNN	80.03	-	-	-
S. Laraba., 2017 [70]	Motion Sequences Maps	CNN	74.27	83.33	-	-
V.-N. Hoang., 2019 [53]	X-Z channel Maps, Velocity Maps	CNN-LSTM	76.8	-	-	-
Our Proposed	Radial basis kernel on joint distances	MRCNNSA	87.67	88.13	90.86	87.34
Our Proposed	Part based Radial basis kernel on joint distances MRCNNSA	MRCNNSA	92.08	92.56	95.43	91.71

increasing the dynamic range of pixels in the colour coded 3D joint features would certainly improve the classifiers performance.

On the other hand, it can further be improved by attention-based learning systems. Finally, from the below comparison tables, we value the proposed method at a 6.24% higher than the current state – of – the – arts with respect to OA metrics. Now, the ultimate question is regarding the placement of multi resolution attention network placement. Till now it mixed at resolution 32 into the mid classification steam in MECNNSA. Does it get better or worse if its position gets shifted to 64 or 16. This study is presented as an ablation study in the final section of this work.

G. ABLATION EXPERIMENTS ON ATTENTION MODULE'S LOCATION AND NUMBER

The MRCNNSA is built on a spatial attention at a specific resolution. Till now it has been projected as 32 which has been giving good results. Its possible that changing the location of the multi resolution spatial attention block forward or backward can impact the overall performance of the classifier. Moreover, it is interesting to evaluate the model's performance with the increasing number of multi resolution spatial attention blocks in parallel across the backbone classifier. Consequently, this ablation study focuses on estimating the performance of the backbone architecture based on the requirement of the number of attention blocks and their position.

1) THE EFFECT OF LOCATION AND RESOLUTION ON MRCNNSA PERFORMANCE

Illustrations on the results obtained in the section IV-F concludes that the proposed multi resolution based spatial attention provides features that boosts recognition accuracy of the classifier. It would be interesting to identify the resolution greater or lesser than 32 at a different location in the backbone R_3 stream of MRCNNSA of Figure.8. Experiments were conducted by selecting various resolutions in the fusion network of Figure.9 and the reaction results are captured using OA metric on KL3DISL dataset. All the mapped features were used for experimentation with three backbone learning architectures shown in Table.8. The primary evidence on feature maps shows that JMAMs are better. Secondly, the resolution other than 32 has shown a downfall in OA of the backbone even for standard models like ResNet50 and VGG-16. The smaller resolutions 16 and 8 were unable to capture the full dynamic range of pixels in the colour coded feature maps. Increasing the image resolution to 64 and 128 for fusion has improved the receptive

TABLE 8. Ablation experiments on MRCNNSA with JMAMs for spatial attention location selection strategies.

Features Considered/ Backbone Networks	Location of SA@Resolution	No Attention	JDMs	JAMs	JADMs	JVMs	JTM	JPM	JMAMs	jg2gMDMs
MRCNNSA	8	0.5987	0.6585	0.7696	0.6856	0.7454	0.6736	0.7204	0.8461	0.8247
	16	0.5567	0.7065	0.7218	0.6446	0.6822	0.6264	0.6574	0.8234	0.8659
	32	0.6103	0.7147	0.7418	0.7725	0.7597	0.6754	0.7569	0.8734	0.9048
	64	0.5812	0.6601	0.7194	0.7358	0.5706	0.6552	0.6625	0.8584	0.8172
	128	0.5737	0.6912	0.7303	0.7114	0.598	0.6327	0.7166	0.8168	0.8635
ResNet50	8	0.5109	0.5795	0.6654	0.7189	0.6508	0.5749	0.5967	0.7822	0.8267
	16	0.4795	0.6486	0.7031	0.7009	0.6954	0.5395	0.7287	0.7531	0.8049
	32	0.5564	0.6951	0.7204	0.7349	0.7283	0.6128	0.7218	0.8481	0.8817
	64	0.4781	0.5943	0.6719	0.7132	0.6517	0.5236	0.7298	0.8451	0.8121
	128	0.5214	0.6668	0.7041	0.7348	0.7061	0.5809	0.6481	0.8381	0.8023
VGG-16	8	0.3864	0.4719	0.4852	0.4969	0.4854	0.4348	0.4888	0.6122	0.6952
	16	0.4112	0.4732	0.4708	0.5238	0.4812	0.4508	0.4918	0.6032	0.6427
	32	0.4681	0.5682	0.5928	0.6018	0.5906	0.5096	0.5976	0.6942	0.7468
	64	0.4107	0.4659	0.4704	0.5364	0.4886	0.4032	0.4844	0.5628	0.6321
	128	0.4119	0.4548	0.4708	0.5132	0.5278	0.4424	0.4888	0.6212	0.6436

TABLE 9. Effective attention fusion strategies for maximizing OA of MRCNNSA on KL3DISL dataset.

Features Considered	Location of SA@Resolution	Number of Attention Modules	JDMs	JAMs	JADMs	JVMs	JTM	JPM	JMAMs
MRCNNSA	128-64	2	0.6474	0.6719	0.6998	0.6882	0.6118	0.6856	0.7912
	128-64-32	3	0.7147	0.7418	0.7725	0.7597	0.6754	0.7569	0.8734
	128-64-32-16	4	0.7004	0.7270	0.7571	0.7445	0.6619	0.7418	0.8559
	128-64-32-16-8	5	0.6761	0.7017	0.7308	0.7187	0.6389	0.7160	0.8262
	128-64-32-8	4	0.6864	0.7124	0.7419	0.7296	0.6687	0.7269	0.8388
	128-64-16-8	4	0.6707	0.6961	0.7249	0.7129	0.6338	0.7103	0.8196
	64-32-16-8	4	0.6633	0.6885	0.7170	0.7051	0.6268	0.7025	0.8106
	128-64-32-8	4	0.6829	0.7138	0.7381	0.7259	0.6453	0.7232	0.8345
	32-16-8	3	0.6394	0.6637	0.6911	0.6797	0.6043	0.6772	0.7814
	64-32-16	3	0.6494	0.6740	0.7019	0.6903	0.6137	0.6877	0.7936
	16-8	2	0.5078	0.5270	0.5489	0.5398	0.4799	0.5378	0.6205
	32-8	2	0.5485	0.5693	0.5929	0.5830	0.5183	0.5809	0.6703
	64-8	2	0.5879	0.6102	0.6355	0.6250	0.5556	0.6227	0.7185
	128-8	2	0.6196	0.6431	0.6697	0.6586	0.5855	0.6562	0.7572

fields of the layers, but this has penalized the OA due to poor encapsulation of local pixel patterns in the maps. However, most of the maps and backbones have recorded highest possible OA at a fusion resolution of 32 by punishing the maximum similarity classes and thus improving their discriminating abilities. At this resolution, both the low- and high-resolution patterns are preserved during fusion process. Obviously, the results show a dip in OA of around 18% in any of the backbone networks when the there is no attention. Attention networks were extensively compared in Table.2 and the metrics show that multi resolution attention maps enhance OA of the moderately designed classifiers.

2) THE IMPACT OF NUMBER OF MULTI RESOLUTION ATTENTION NETWORKS ON CLASSIFIER OA

Undoubtedly, the mutli resolution fusion maps at 32 have improved the OA of the 3D sign language recognition. The improvement is around 7 to 8% from the previous methods. It would be interesting to verify the multi resolution fusions at more than one location. In the previous study, we selected only one location and resolution for fusion in the backbone networks. Here, we select multiple locations and resolutions for fusion. For example, resolution 64 after the 2^{nd} convolutional layer along with the regular fusion across 3^{rd} layer and a resolution of 16 across the 4^{th} in the backbone network. Accordingly, all possible combinations were selected during training and testing of the backbone classifiers. The results are presented in Table.9. The listed values in the Table.9 concludes that the early fusions strategies across the first two or middle three convolutional layers has greater impact on the overall performance of the classifier when compared to other combinations. This is due to remarkable structural integrity in pixel patterns that is prominent at higher and middle image resolutions. Mid-level feature fusion at multiple locations has certainly increased the OA of the classifier but has also increased the computation time and complexity. Arguably, the OA increase is not highly significant when compared to the increase in complexity of the classifier. The experiments were executed on NVIDIA 8GB GTX1070 graphics processor with 16GB DDR4 RAM. The results show the effectiveness of using the proposed joint

movement affinity maps (JMAMs) and a spatial resolution based attention model for 3D skeletal sign (action) language recognition tasks.

V. CONCLUSION

The current generation of spatiotemporal 3D Joint features for action or sign language recognition were flaged ineffective due to inconsistent motion pixel distributions. Interestingly this work discovers JMAMs for mapping 3D joint motions to color coded features trained on MRCNNSA network. The JMAMs guarantees a positive proportionality mapping of 3D motion information into color coded images which have non uniform hight to width ratios. Traditional single resolution transformations on color coded maps failed to classify highly correlated features between labels. The MRCNNSA develops consistent features for classification across multiple resolutions that represent both local and global pixel variations efficiently. The joint group to group motion directed JMAMs (jg2gMDMs) were also trained and tested for SLR on MRCNNSA model. The KL3DSL trained MRCNNSA classifies 3D signs with an OA of 87.34% on JMAMs and 91.71% jg2gMDMs, which is ±5% higher than previous works respectively. The results on 3D human action datasets have proved that the JMAMs have high 3D motion consistency for transferring them in to color codings. Moreover, the MRCNNSA cohesiveness with JMAMs has improved the impact of 3D sign language recognition system. In future, the proposed JMAMs computed from estimated pose can be correlated with our 3D model poses, which are then inferenced on trained MRCNNSA for real time sign language translator.

REFERENCES

[1] F. Wen, Z. Zhang, T. He, and C. Lee, "AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove," *Nature Commun.*, vol. 12, no. 1, pp. 1–13, Sep. 2021.

[2] J. Fenlon, A. Schembri, R. Rentelis, D. Vinson, and K. Cormier, "Using conversational data to determine lexical frequency in British sign language: The influence of text type," *Lingua*, vol. 143, pp. 187–202, May 2014.

[3] R. Gupta and A. Kumar, "Indian sign language recognition using wearable sensors and multi-label classification," *Comput. Electr. Eng.*, vol. 90, Mar. 2021, Art. no. 106898.

[4] D. A. Kumar, P. V. V. Kishore, A. S. C. S. Sastry, and P. R. G. Swamy, "Selfie continuous sign language recognition using neural network," in *Proc. IEEE Annu. India Conf. (INDICON)*, Dec. 2016, pp. 1–6.

[5] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113794.

[6] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 7957–7968, Jun. 2020.

[7] K. Nimisha and A. Jacob, "A brief review of the recent trends in sign language recognition," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 186–190.

[8] M. Mustafa, "RETRACTED ARTICLE: A study on Arabic sign language recognition for differently abled using advanced machine learning classifiers," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 3, pp. 4101–4115, Mar. 2021.

[9] G. A. Rao, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry, "Deep convolutional neural networks for sign language recognition," in *Proc. Conf. Signal Process. Commun. Eng. Syst. (SPACES)*, Jan. 2018, pp. 194–197.

[10] A. Chaikaew, K. Somkuan, and T. Yuyen, "Thai sign language recognition: An application of deep neural network," in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng.*, Mar. 2021, pp. 128–131.

[11] K. Bantupalli and Y. Xie, "American sign language recognition using deep learning and computer vision," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 4896–4899.

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

[13] J. Forster, C. Oberdörfer, O. Koller, and H. Ney, "Modality combination techniques for continuous sign language recognition," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.* Cham, Switzerland: Springer, 2013, pp. 89–99.

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

[15] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1311–1325, Dec. 2018.

[16] A. Sadeghzadeh and M. B. Islam, "Triplet loss-based convolutional neural network for static sign language recognition," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, Sep. 2022, pp. 1–6.

[17] T. Raghuveera, R. Deepthi, R. Mangalashri, and R. Akshaya, "A depth-based Indian sign language recognition using Microsoft Kinect," *Sādhanā*, vol. 45, no. 1, pp. 1–13, Dec. 2020.

[18] E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, and D. A. Kumar, "3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2—Stream CNN," *Neurocomputing*, vol. 372, pp. 40–54, Jan. 2020.

[19] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3D human-skeleton sequences for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3054–3062.

[20] J. C. Nuñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognit.*, vol. 76, pp. 80–94, Apr. 2018.

[21] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017.

[22] M. Naveenkumar and S. Domnic, "Spatio temporal joint distance maps for skeleton-based action recognition using convolutional neural networks," *Int. J. Image Graph.*, vol. 21, no. 5, Dec. 2021, Art. no. 2140001.

[23] T. K. K. Maddala, P. V. V. Kishore, K. K. Eepuri, and A. K. Dande, "YogaNet: 3-D yoga asana recognition using joint angular displacement maps with ConvNets," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2492–2503, Oct. 2019.

[24] M. T. K. Kumar, P. V. V. Kishore, B. T. P. Madhav, D. A. Kumar, N. S. Kala, K. P. K. Rao, and B. Prasad, "Can skeletal joint positional ordering influence action recognition on spectrally graded CNNs: A perspective on achieving joint order independent learning," *IEEE Access*, vol. 9, pp. 139611–139626, 2021.

[25] M. T. K. Kumar, P. V. V. Kishore, and M. V. D. Prasad, "CNN-LSTM hybrid model based human action recognition with skeletal representation using joint movements based energy maps," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 7, pp. 3502–3508, Jul. 2020.

[26] K. Thakkar and P. J. Narayanan, "Part-based graph convolutional network for action recognition," 2018, *arXiv:1809.04983*.

[27] Y. Qin, L. Mo, C. Li, and J. Luo, "Skeleton-based action recognition by part-aware graph convolutional networks," *Vis. Comput.*, vol. 36, no. 3, pp. 621–631, Mar. 2020.

[28] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[29] CMU. (2013). *CMU Graphics Lab Motion Capture Database*. [Online]. Available: http://mocap.cs.cmu.edu/

[30] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database HDM05," Tech. Rep., 2007.

[31] D. A. Kumar, A. S. C. S. Sastry, P. V. V. Kishore, and E. K. Kumar, "Indian sign language recognition using graph matching on 3D motion captured signs," *Multimedia Tools Appl.*, vol. 77, no. 24, pp. 32063–32091, Dec. 2018.

[32] P. V. V. Kishore, D. A. Kumar, and M. Manikanta, "Continuous sign language recognition from tracking and shape features using fuzzy inference engine," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 2165–2170.

[33] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 2257–2264.

[34] V. Adithya, P. R. Vinod, and U. Gopalakrishnan, "Artificial neural network based method for Indian sign language recognition," in *Proc. IEEE Conf. Inf. Commun. Technol.*, Apr. 2013, pp. 1080–1085.

[35] M. Suneetha, M. V. D. Prasad, and P. V. V. Kishore, "Multi-view motion modelled deep attention networks (M2DA-Net) for video based sign language recognition," *J. Vis. Commun. Image Represent.*, vol. 78, Jul. 2021, Art. no. 103161.

[36] H. R. Vaezi Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13286–13296.

[37] E. Rajalakshmi, R. Elakkiya, A. L. Prikhodko, M. G. Grif, M. A. Bakaev, J. R. Saini, K. Kotecha, and V. Subramaniyaswamy, "Static and dynamic isolated Indian and Russian sign language recognition with spatial and temporal feature detection using hybrid neural network," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 1, pp. 1–23, Jan. 2023.

[38] B. Natarajan, E. Rajalakshmi, R. Elakkiya, K. Kotecha, A. Abraham, L. A. Gabralla, and V. Subramaniyaswamy, "Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation," *IEEE Access*, vol. 10, pp. 104358–104374, 2022.

[39] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent Kinect-based action recognition algorithms," *IEEE Trans. Image Process.*, vol. 29, pp. 15–28, 2020.

[40] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the Kinect," in *Proc. 13th Int. Conf. Multimodal Interface*, Nov. 2011, pp. 279–286.

[41] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3D skeleton-based action recognition using learning method," 2020, *arXiv:2002.05907*.

[42] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3010–3022, Jul. 2016.

[43] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 579–583.

[44] E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, D. A. Kumar, and A. S. C. S. Sastry, "Three-dimensional sign language recognition with angular velocity maps and connived feature ResNet," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1860–1864, Dec. 2018.

[45] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[46] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2330–2343, Sep. 2018.

[47] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowl.-Based Syst.*, vol. 158, pp. 43–53, Oct. 2018.

[48] Z. Wang, B. Wang, H. Liu, and Z. Kong, "Recurrent convolutional networks based intention recognition for human–robot collaboration tasks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1675–1680.

[49] J. Dong, Y. Gao, H. J. Lee, H. Zhou, Y. Yao, Z. Fang, and B. Huang, "Action recognition based on the fusion of graph convolutional networks with high order features," *Appl. Sci.*, vol. 10, no. 4, p. 1482, Feb. 2020.

[50] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Ongoing human action recognition with motion capture," *Pattern Recognit.*, vol. 47, no. 1, pp. 238–247, Jan. 2014.

[51] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Volume 1: Long Papers)*, 2015.

[52] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4382–4394, Sep. 2018.

[53] V.-N. Hoang, T.-L. Le, T.-H. Tran, and V.-T. Nguyen, "3D skeleton-based action recognition with convolutional neural networks," in *Proc. Int. Conf. Multimedia Anal. Pattern Recognit. (MAPR)*, May 2019, pp. 1–6.

[54] S. Huang, C. Mao, J. Tao, and Z. Ye, "A novel Chinese sign language recognition method based on keyframe-centered clips," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 442–446, Mar. 2018.

[55] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image for human action recognition," *Pattern Recognit.*, vol. 45, no. 7, pp. 2562–2572, Jul. 2012.

[56] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Part-level graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11045–11052.

[57] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 915–922.

[58] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 392–404, Feb. 2018.

[59] P. V. V. Kishore, D. A. Kumar, A. S. C. S. Sastry, and E. K. Kumar, "Motionlets matching with adaptive kernels for 3-D Indian sign language recognition," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3327–3337, Apr. 2018.

[60] W. Chan, Z. Tian, and Y. Wu, "GAS-GCN: Gated action-specific graph convolutional networks for skeleton-based action recognition," *Sensors*, vol. 20, no. 12, p. 3499, Jun. 2020.

[61] J. Zhu, W. Zou, Z. Zhu, and Y. Hu, "Convolutional relation network for skeleton-based action recognition," *Neurocomputing*, vol. 370, pp. 109–117, Dec. 2019.

[62] W. Ding, X. Li, G. Li, and Y. Wei, "Global relational reasoning with spatial temporal graph interaction networks for skeleton-based action recognition," *Signal Process., Image Commun.*, vol. 83, Apr. 2020, Art. no. 115776.

[63] X. Chen and M. Koskela, "Skeleton-based action recognition with extreme learning machines," *Neurocomputing*, vol. 149, pp. 387–396, Feb. 2015.

[64] L. Wang, X. Zhao, and Y. Liu, "Skeleton feature fusion based on multi-stream LSTM for action recognition," *IEEE Access*, vol. 6, pp. 50788–50800, 2018.

[65] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1044–1048, Jul. 2018.

[66] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[67] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," in *Proc. CVPR Workshops*, 2019, pp. 10–19.

[68] J. Ren, N. H. Reyes, A. L. C. Barczak, C. Scogings, and M. Liu, "An investigation of skeleton-based optical flow-guided features for 3D action recognition using a multi-stream CNN model," in *Proc. IEEE 3rd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2018, pp. 199–203.

[69] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 601–604.

[70] S. Laraba, J. Tilmanne, and T. Dutoit, "Leveraging pre-trained CNN models for skeleton-based action recognition," in *Proc. Int. Conf. Comput. Vis. Syst.* Cham, Switzerland: Springer, 2019, pp. 612–626.

[71] K. V. Prasad, P. V. V. Kishore, and O. S. Rao, "Skeleton based view invariant human action recognition using convolutional neural networks," *Int. J. Recent Technol. Eng.*, vol. 9, no. 2, pp. 4860–4867, Jul. 2019.

[72] C. Caetano, F. Brémond, and W. R. Schwartz, "Skeleton image representation for 3D action recognition based on tree structure and reference joints," in *Proc. 32nd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2019, pp. 16–23.

[73] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.