Computers, Materials & Continua
DOI: 10.32604/cmc.2023.031924
Article

Tech Science Press

A Novel Action Transformer Network for Hybrid Multimodal Sign Language Recognition

Sameena Javaid* **and Safdar Rizvi**

Department of Computer Sciences, School of Engineering and Applied Sciences, Bahria University, Karachi Campus, Karachi, Pakistan
*Corresponding Author: Sameena Javaid. Email: sameenajaved.bukc@bahria.edu.pk
Received: 30 April 2022; Accepted: 22 June 2022

Abstract: Sign language fills the communication gap for people with hearing and speaking ailments. It includes both visual modalities, manual gestures consisting of movements of hands, and non-manual gestures incorporating body movements including head, facial expressions, eyes, shoulder shrugging, etc. Previously both gestures have been detected; identifying separately may have better accuracy, but much communicational information is lost. A proper sign language mechanism is needed to detect manual and non-manual gestures to convey the appropriate detailed message to others. Our novel proposed system contributes as Sign Language Action Transformer Network (SLATN), localizing hand, body, and facial gestures in video sequences. Here we are expending a Transformer-style structural design as a "base network" to extract features from a spatiotemporal domain. The model impulsively learns to track individual persons and their action context in multiple frames. Furthermore, a "head network" emphasizes hand movement and facial expression simultaneously, which is often crucial to understanding sign language, using its attention mechanism for creating tight bounding boxes around classified gestures. The model's work is later compared with the traditional identification methods of activity recognition. It not only works faster but achieves better accuracy as well. The model achieves overall 82.66% testing accuracy with a very considerable performance of computation with 94.13 Giga-Floating Point Operations per Second (G-FLOPS). Another contribution is a newly created dataset of Pakistan Sign Language for Manual and Non-Manual (PkSLMNM) gestures.

Keywords: Sign language; gesture recognition; manual signs; non-manual signs; action transformer network

1 Introduction

People who are deaf and mute rely heavily on sign language for communication. People move their hands to communicate as a nonverbal way of expressing thoughts, needs, and messages. The World Health Organization (WHO) declared it deafened 430 million individuals [1]. Deaf and mutes communicate with regular people with various signs from sign language. Sign language includes both

manual and non-manual gestures. In sign language, visual modalities are manual and non-manual gestures. Manual gestures utilize hands and signs only, whereas non-manual gestures incorporate body movements, including head, shoulder shrugging, facial expression, etc. A common way of communication of including deaf, mute, and ordinary people is a facial expression [2]. It is a rational mental reaction towards any particular object that can feel like a strong feeling and is an essential aspect of effective communication, usually followed by physiological responses. Nonverbal communication takes the form of facial expressions and hand gestures as it expresses humans' feeling a lot quicker than verbal communication and is more effective. Neutral, happy, sad, angry, disgust, surprise, and fear are a few universal emotional expressions. A person's emotion can be analyzed in various stages as it defines a person's current state, mood, and feelings. According to Mehrabian, a Psychologist, emotion is conveyed 7 percent through language, 38 percent through voice, and 55 percent through facial expression [3].

A rational mental reaction towards any particular object, which can feel like a strong feeling and is an essential aspect of effective communication, is an emotion, usually followed by physiological responses. One of the most critical impacts on our life is emotion, as they show our physical and behavioral changes. Nonverbal communication, including audio, videos, photographs, and images, is difficult to recognize. To communicate with one another in various ways, humans can do so through facial expression, body language, speech, and body gestures [2]. The human brain can anticipate the other person's emotion depending upon the situation based on the current mood standing in front of them. Nonverbal communication takes the form of facial expressions as it expresses human feelings a lot quicker than verbal communication and is more effective. A human's face spreads emotional information in terms of facial expressions, including eyes, brows, mouth, and cheeks. Neutral, happy, sad, angry, disgust, surprise, and fear are a few universal emotional expressions.

Currently, automatic Sign Language (SL) analysis and recognition are considered by many researchers, precisely intuitive SL interpretation. Two convincing ways to conceive such systems are: using special devices like Leap Motion Controller (LMC), Kinect sensors, or other motion sensors [4,5]. Another way is vision-based automatic inspection and analysis [6], where researchers perform machine learning and vision-based deep learning to understand Sign Language (SL) [7]. Techniques based on some devices did not arouse well, requiring signers to wear expensive additional devices. Therefore, vision-based techniques are more adequate and require fewer devices and resources.

Machine learning and deep learning convolutional neural networks can successfully recognize various alphabets and numbers of different sign languages and human facials such as good, bad, disgusted, anger, happiness, sad, surprise, fear, etc. in both static and dynamic images in real-time but unfortunately, very few contributions are made to combining sign language and facial expression of people considering deaf [8]. Working on only one problem simultaneously creates a loss of useful information. It is also observed that signers' facial expressions and body posture frequently change while performing any sign to deliver the actual meaning and clear sense of the gesture concerning language. Similarly, a single sign may correspond to several facial gestures, and a single facial expression can combine with many signs. The point of the proposed strategy is to distinguish and perceive sign language, including body gestures of hands, arms, and facial expressions. These feelings involve profound learning with the most extreme exactness in a restricted chance to rival conventional techniques. The main contributions of the proposed work are as follows:

- Our first contribution is a novel Sign Language Action Transformer Network (SLATN) architecture to interpret sign language with manual and non-manual features.

- Our second contribution is creating a local database of Pakistan Sign Language (PSL) using seven basic expressions representing seven various adjectives in language, naming Pakistan Sign Language using Manual and Non-Manual (PSLMNM). To the best of our knowledge, this is the first dataset of PSL having manual and non-manual gestures combined.

SLATN can transform the performed action effectively to a different view at the abrasive level. But the main limitation of the current architecture is that sometimes the higher appearance features are missing in the recorded video leading to motion blur. In the future, the availability of resources will make it possible to improve the quality but apprehend delicate appearance and motion information using some memory constrictions environment.

The following is the order in which the paper should be interpreted. Section 2 outlines state-of-the-art literature on activity detection and recognition using deep learning and computer vision. The proposed technique for detecting the signs of deaf and mute people is presented in Section 3. Section 4 spoke about how we generated our dataset, analyzed the results, and compared them with existing work. Finally, Section 5 brings the study debate to a brief conclusion.

2 Literature Review

Machine learning and deep learning networks are more widely used as they evolve regularly. These systems and networks make life much easier for us as now they've become an indispensable part of our modern lives [9]. Deep learning sign language gesture and facial recognition systems capture different human emotions and messages and have overcome traditional methods in terms of accuracy and speed. Many researchers have proposed their work in recognizing and detecting signs of varying sign languages and human emotions individually and successfully for normal and people with hearing and speaking disabilities.

It is difficult for people with hearing disabilities to communicate. Also, this disability affects their understanding of their emotions as well. Yuan Tao et al. designed an application to read and recognize facial expressions specifically for deaf and mute people [10], which helps to cross-language, emotional, and regional barriers for deaf and mute people. The designed app was tested on 630 images of gestures and scored an incredible accuracy of 94.22%. Another research study proposed a model by studying various emotions and electroencephalograph (EEG) signals of physically impaired people and autistic children through convolutional neural networks with Long Short Term Memory (LSTM) classifiers in Kuwait [11]. The aim was to study the prominent rising need to recognize different facial expressions and emotions utilizing virtual markers in an optical flow model that worked efficiently and achieved maximum accuracy of 99.81%.

Similarly, in another study, emotions are recognized using body gestures. The proposed model extracted features from the input video from an online, freely accessible dataset through the hashing method and used convolutional LSTM to utilize the extracted sequential information [12]. Working on the same dataset, [13] suggested a model highlighting the same problem: identifying emotions using upper body movements and facial expressions. The convolutional neural network (CNN) extracts the features while LSTM utilizes the sequential information and achieves 94.41% accuracy. It is also recognized for independent sign language, comprising a 3D body, hands, and facial expressions. For this purpose, SMPL-X extracts features of body shape, face, and writing using a single image [14]. The 3D reconstructed output of the 3D model is utilized for SLR resulting in higher accuracy than raw RGB images.

Sign language acts as an indispensable source of communication for people with hearing and speech disabilities, but detecting sign in a continuous video face many challenges in terms of accuracy and performance. Previously, most works were done on seeing any modality with static images. Few researchers proposed a method to detect sign language in a continuous video, including gestures of both hands, head pose, and eye gaze [15]. Features from the continuous video were extracted by using Hidden Markov Model (HMM) and classified gestures through Independent Bayesian Classifier Model (IBCC) with 93% of good accuracy. Another researcher used an auto-encoder convolutional network for a large-scale RGB video to design a novel 3D Gesture Segmentation Network (3D GS-Net) for developing a communicational environment for deaf people utilizing hands, body, and facial expressions [16]. Compared with other state-of-the-art language recognition systems, the proposed model was tested using Moroccan sign language (MoSL) and performed better.

In summary, it is stated that several researchers studied manual or non-manual gesture interpretation in sign language but working with the correlation between the two is rare in the literature, along with better accuracy and speed using video processing or dynamic sign language interpretation considering full-body gestures. A raw High Definition (HD) video processing, while considering manual and non-manual gestures with maximum accuracy and limited time is still a requirement for Dynamic SLR.

3 Methodology

The section elaborates on the complete design and flow of the model. The proposed model is intended to classify humans and characterize their emotions and gestures at a specific time. It takes video clips as input and generates labeled bounding boxes around people doing activities appearing in the video. The model has two separate networks, i.e., base and head like Faster Region-based Convolutional Neural Network (F-RCNN). The base expends a 3Dimensional-Convolution structure to extract features, and region bounds are generated for the individuals present in the input video. The head later utilizes these characteristics to predict bounding boxes with better accuracy, corresponding to each Region Proposal (RP). The head uses the features representation created by the base network using Region of Interest (RoI) Pooling to predict the class label and regressor bounding boxes. These boxes are categorized as action classes with the background. The RPN proposal is regressed to a 4D vector to create strict bounding boxes around people. The base and the head networks are described below.

3.1 Base Network

3.1.1 I3D Model

The conventional deep learning models utilized a mono convolutional kernel, and before the feature set is generated, this convolution kernel processes its input. Different kernels are used to analyze the sprite separately in the Inception module. Although the resulting features are different, a group of complementary features creates a subset of densely scattered features. Consequently, after passing various convolutional layers, unnecessary data is undermined. The Inflated 3-Dimension (I3D) module comprises three inception layers and two convolutional pooling layers.

The I3D module is inherited after GoogLeNet's Inception network, but different sizes of convolutional kernels extract the features. In GoogLeNet [17], one convolution is performed on the yield of each prior layer which is succeeded by an activation function. Additional nonlinear features are joined by adding two three-dimensional convolutions. The inception module has four groups of input data, whereas each group consists of one or more convolution and pooling operations. Finally, different

convolution kernels of various sizes are spliced together. The I3D model has a convolution operation for every dataset's adjacent feature, which continuously completes the action recognition on frames. A batch regularization module is connected to the system to accelerate the training process. A higher learning rate can be used since the initialization does not affect the model. To enhance the depth of the model, I3D has eight convolutional and four pooling layers. The kernel has a step size of $1 \times 1 \times 1$ and a pixel size of $3 \times 3 \times 3$ in each convolutional layer. The numbers of filters are multiples of 64 along with each convolutional layer containing batch regularization along with an activation function with a pooling layer, respectively. The pooling layer and the step size of the kernels in conv3, conv4, and conv5 are $1 \times 2 \times 2$. The remaining pooling layers have the same step size and kernel. There is only spatial pooling in the first convolution layer, while the second, fourth and sixth convolution layers have spatial-temporal pooling layers used. The yield size from convolutional layers is limited to $1/4$ and $1/2$ in space and time regions because of the pooling layers. As a result, I3D is more suitable for LSTM with spatial-temporal features.

3.1.2 I3D ShuffleNet

In conventional I3D, convolutional operations are performed by two kernels of equal $5 \times 5 \times 5$ size to extract features but at the same time cause excessive computational power. These kernels can be modified by following specific dimensional rules, which will convolve images after learning different kernel combinations. Fig. 1 shows kernels of $3 \times 3 \times 3$ size, which replace the $5 \times 5 \times 5$ convolutional kernel. In this replacement, almost 30% of trainable parameters are reduced.

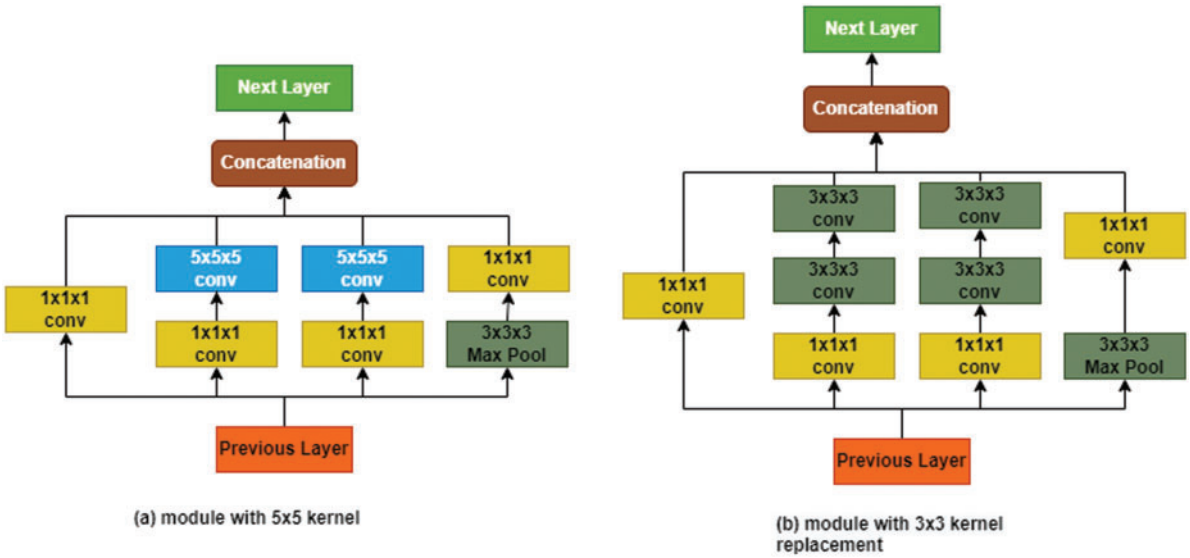


Figure 1: Inception module with 5×5 and 3×3 convolution kernel

3.1.3 Channel Shuffling

The channel shuffle is made by using the concept of shuffleNet. As a deep learning network, Face + + presented the shuffleNet, a CNN model that is highly efficient and robust. It is primarily designed to use with communication systems such as robots, drones, and phones and hence strives as the finest network in terms of accuracy and restricting computational power. The primary task of shuffleNet, which immensely reduces the number of computations while keeping the accuracy high, is

the channel shuffle and element-wise group convolution. The basic I3D, on the other hand, only uses group convolution, which appears as its major disadvantage because its output channel is generated by using the small chunks of the input channel. For shuffling, pixel-level group convolution is created, which decreases the complexity of computation generated due to the convolution operation. Group convolution hampers the exchange of data between the channels, resulting in loss of feature extraction. Channel shuffling resolves this issue of exchanging information between the channels. Fig. 2 represents the channel shuffling in a descriptive way.

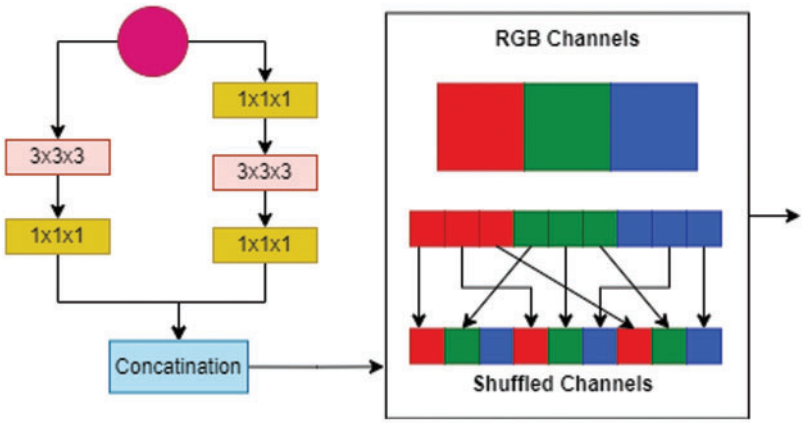


Figure 2: Channel shuffling

The process of splitting channels is introduced. For the feature maps, the c channel splits into two sections. The first section consists of c-c' channels, while the second has only c' channels. One part has an equal number of three convolutional channels, while the other is fixed to limit the amount of shuffle. Channel segmentation is the second part of splitting two groups, and convolutions of two 1×1 are not joined. To maintain the number of channels constant, the features of the two parts are merged to ensure that the information of these two parts engages. To generate the residual block, medians and point-to-point convolution are mixed. To enhance the flow of data through various groups on the main feature of the bottleneck branch, the shuffleNet operation is succeeded by point-to-point grouping convolution. The computation is also reduced by adding small depth split table branch pixel convolution of sizes $3 \times 3 \times 3$ and $1 \times 1 \times 1$ after point to point grouping convolution. A layer of max-pooling is introduced to replace the pixel by pixel summing operation, which can cause an increase in computation by expanding the dimension of the channel.

3.1.4 Structure

After incorporating time information, for I3D shuffleNet, the 2D convolution is enlarged to 3D shuffleNet. In the proposed architecture SLATN, the instant normalization layer is introduced instead of typical batch normalization. The input image features are merged using processing of the various inception convolution while the input of the channel shuffle is coming from the output of the inception network, with 50 percent of the feature maps being inputted directly into the following network. The shuffle operation works after the 6th layer inception module; the resulting output is merged with the 9th Inception module. The exact terms for reusing features are used in DenseNet and CondenseNet. Later three channels split the second half and were analyzed independently using channel segmentation. The whole batch needs to be loaded in memory for batch normalization, which requires more memory than instant segmentation. The architecture is shown in Fig. 3.

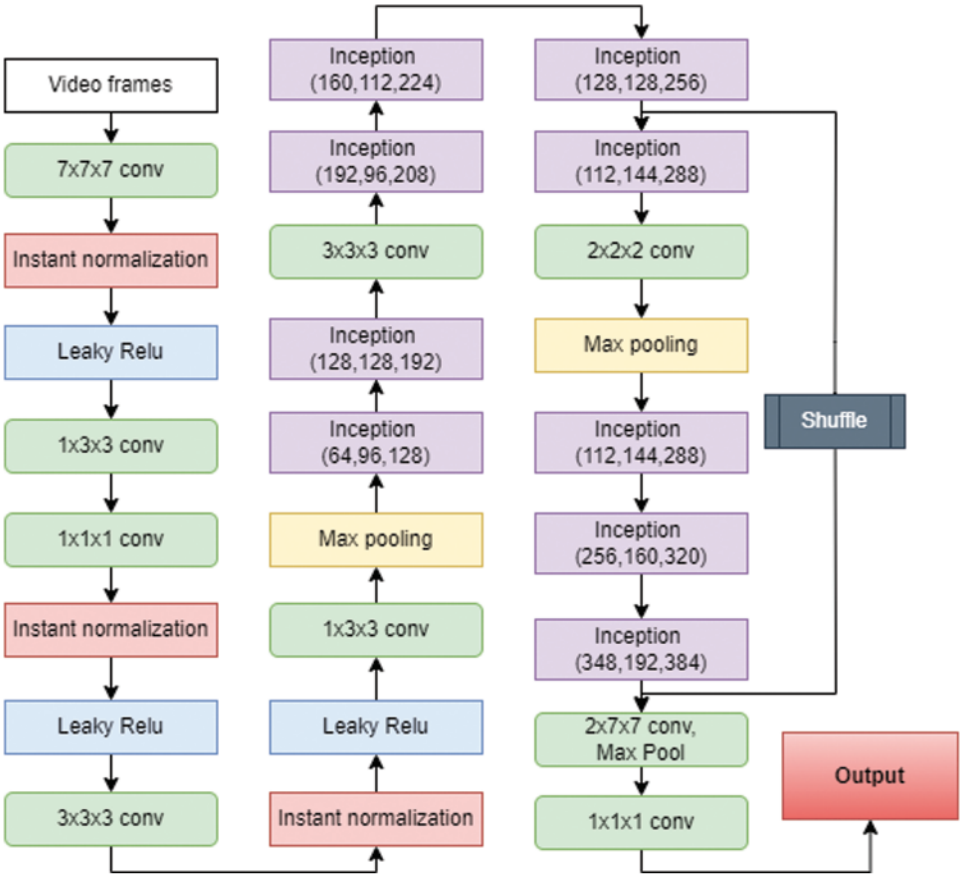


Figure 3: I3D ShuffleNet

3.2 Head Network

3.2.1 Sign Language Action Transformer Framework

As the introduction states, the head of our framework is prompted by and reconfigured from the Transformer architecture [18]. It locates regions to join to use the RPN's user box as a 'query' and categorizes the data over the video to categorize their actions. The proposed architecture of the action transformer head is to replace conventional reoccurring models for seq2seq activities like translators. The essential principle of the prevailing is to calculate self-attention, which is performed by comparing each feature in a series. This is accomplished effectively by avoiding the use of the original features. They are instead utilizing linear projections. Initially, features are mapped to query and memory, represented by (Q) and (K as key & V as value), having low dimensions. Where the result of the query is calculated as the sum of V (all weighted attention), attention weights are acquired from Q and K. The proposed paper had a word as a query that had to get translated in practice; the input ad linear projections as keys and values linear whereas the sequences of the output were generated. The lost information due to a non-convolutional setup, a system of embedding location is added for further details readers can read [18] and [19].

3.2.2 Action Transformer

The proposed model uses a modified action transformer structure to understand videos better. In this case, the values of the query (Q), key (K), and memory are fitted naturally according to the problem: the recognized person is the query, and memory is the video of the person surrounding estimated into values and keys. The unit and the memory analyze the query to generate query vectors. The unit then processes the query and memory, which results in the newly developed "query vector." The supposition that self-attention will assist the "query vector" with classification by adding context from the surroundings in the video. By convolving the yield from various heads at a specific layer, cascade features will be the following query; this unit could be heaped in different layers and head sequences, close to the initial structure [18]. The next layer uses this latest query to pay more attention to context features. Fig. 4 depicts the above setup formfitting in our base network, 'Tx' is the sign language action transformer unit, represented in blue. Further, this unit is explained thoroughly.

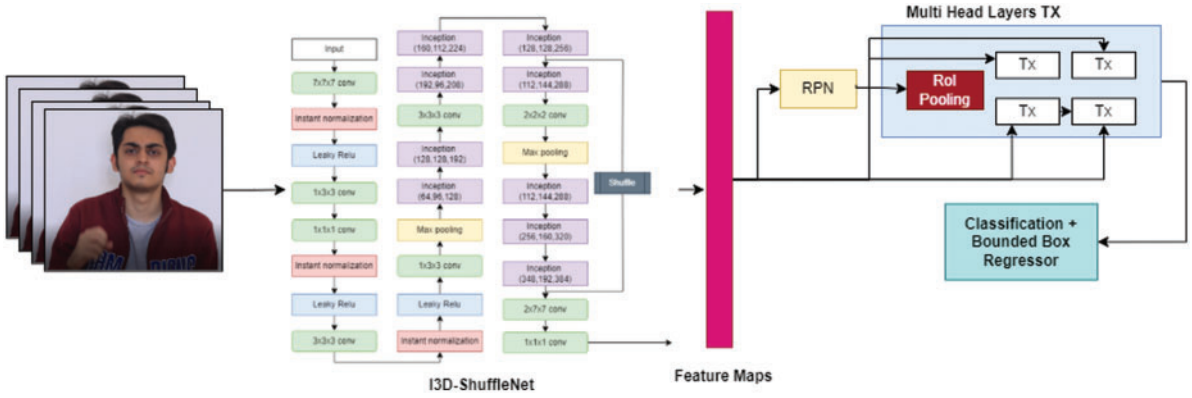


Figure 4: The system architecture of the proposed model

The trunk output gives the initial feature maps for the key and value features, resulting in a shape of T′ H′ W′ D for each.

While performing, query features are extracted from the center clip by ROI pooled features having size 11 D for individual boxes and passed through a linear layer and query pre-processor(QPr). The pooled features from ROI could be in a straight line averaged across the (QPr). However, the individual's spatial layout would be lost. Instead, we combine the derived 7×7 feature map cells into a vector after reducing the dimensionality with a 1×1 convolution. At last, we use a linear layer to limit the dimensionality of this feature map to 128D. This process, known as the HighRes query preprocessing Swish function, is used as an activation to avoid the slow process of the training process around the zero gradients. Compared to the RPN proposal r, the Q-r feature is utilized. The K features are normalized by $\sqrt{}$D; the dot product is used. This operation can be expressed briefly in Eq. (1) [20].

$$a_{xyt}^{(r)} = \frac{Q^r K_{xyt}^T}{\sqrt{D}}; A^r = \sum_{x,y,t} \left[Swish \quad \left(a^{(r)}\right) \right]_{xyt} \quad V_{xyt}. \tag{1}$$

An (r) gets a dropout and is added to the initial features of the query. The query can then pass through a remaining branch that includes a Layer Norm operation, a "Feed-Forward Network" (FFN) implemented as a 2-layer MLP, and dropout. To obtain the revised query (Q″), the concluding feature

is passed through one more Layer Norm. Fig. 4 (Tx unit) depicts the unit mentioned above structure and can be symbolized in the following Eqs. (2) and (3).

$$Q^{(r)'} = Layer\ \ Norm\ \ (Q^{(r)} + Dropout\left(A^{(r)}\right), \tag{2}$$

$$Q^{(r)''} = Layer\ \ Norm\ \ (Q^{(r)'} + Dropout(FFN\left(Q^{(r)'}\right). \tag{3}$$

4 Experiment

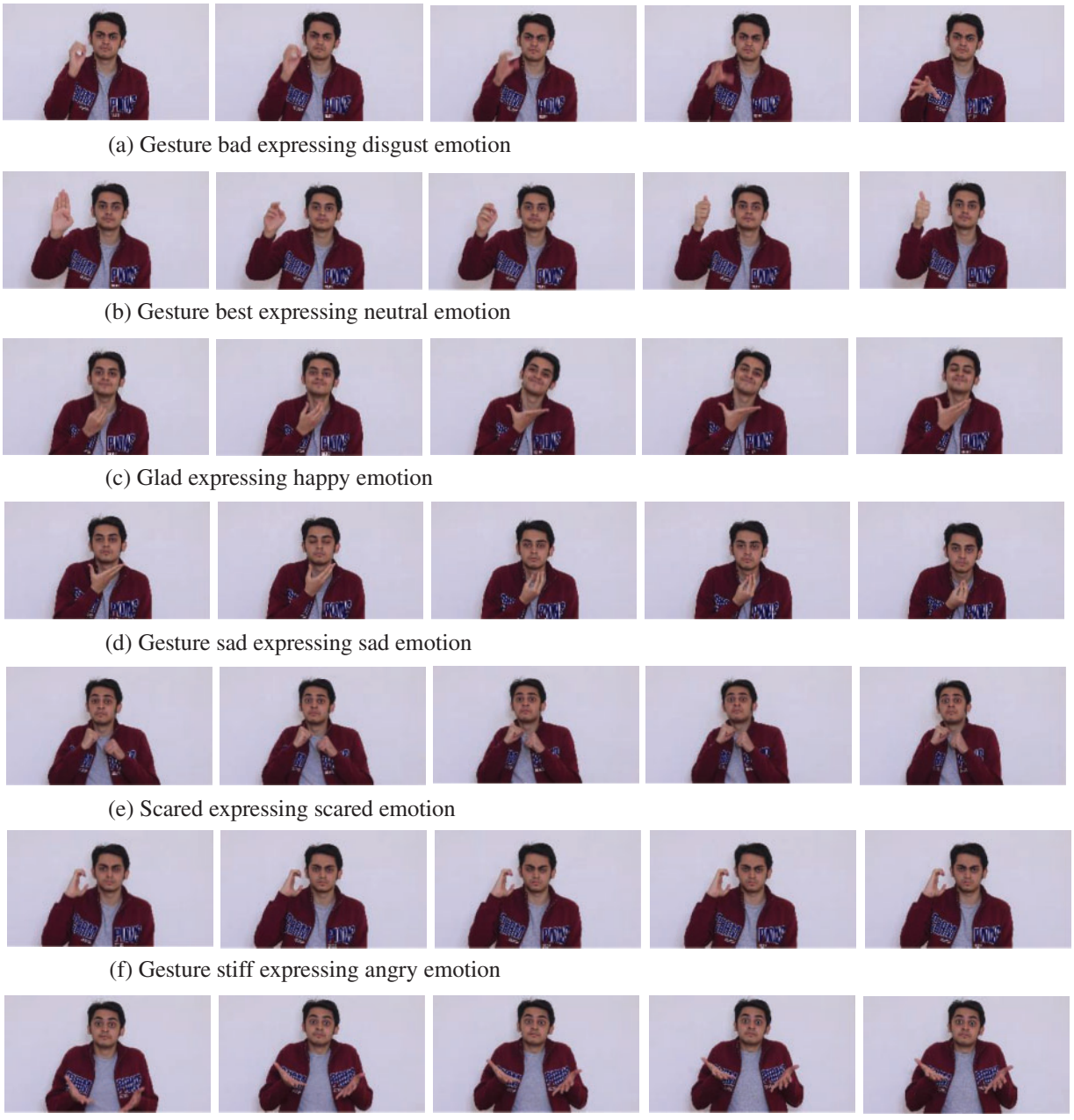
4.1 Dataset

To identify manual and non-manual gestures simultaneously, a dataset of Pakistan Sign Language (PSL), named Pakistan Sign Language Manual and Non-Manual (PkSLMNM), consists of 180 people with an average age group of 20 to 50. Of which 70 are females, and the rest 110 are males. All of the participants have submitted consent to taking part in this study. Furthermore, regarding ethical norms for data collection and usage, the ethical review committee of the School of Engineering and Applied Sciences, Bahria University Karachi Campus, has approved the data collection process and purpose of research under application ERC/ES/CS/005. PSLMNM dataset is solely collected for research purposes, and we ensure the procedure adapted for data collection is not harmful to the participants.

Several researchers have worked on the computational automation of Pakistan Sign Language (PSL), but none were published online for open access. Initially, in the early 20 s Deaf Reach Program by Pakistan Sign Language (PSL) organization created a repository of Static and Dynamic signs of 5000 signs [21,22]. Recently from 2019 to 2022, few researchers [23,24], have taken steps toward publishing some datasets for computer vision, machine learning, and deep learning [25]. These published datasets are collections of static sign language gestures comprised of still images. None of them are casing dynamic sign language gestures with multi-modalities in the form of videos. Also, the scope and variety of the dataset are limited, which results in tradeoffs between the efficiency of the system while at the same time giving flexibility to the system. The primary objective of the current dataset PkSLMNM is to serve as a new benchmark in Pakistan Sign Language with the unique feature of dynamic gestures along with manual and non-manual modalities combined. The lexicon of this current dataset is based on dynamic hand gestures. To the best of our knowledge, no dataset covers the dynamic gestures in the publicly available dataset. The current dataset is also peerless for its vast lexicon, robustness, high recording quality, and unique dynamic hand gesture recognition syntactic property.

The video clip for the input is recorded using an HD camera in .MP4 format to save various facial expressions supported by hand gestures such as bad, best, sad, glad, scared, stiff, and surprised. The individual candidate has to keep a specific emotion for a few seconds, the average duration of all recorded clips is 2 s long, and the average size of the input clip is 3MB. Figs. 5a–5g represents bad, best, sad, glad, scared, stiff, and surprise adjectives of PSL portraying disgust, neutral, sad, happy, scared, angry, and surprise expressions, respectively.

The video clips are then preprocessed into frames at a rate of 25 frames/s to remove any noise which can later affect the classification. The height and width of each frame are 1920×1080. After removing noise, the dataset is passed through the augmentation steps such as flips, contrast, and cropping. The augmented frames are well labeled with all seven different classes of emotions. The preprocessed data is divided into training, testing, and validation sets. Dataset is publically available for usage [26].



(a) Gesture bad expressing disgust emotion

(b) Gesture best expressing neutral emotion

(c) Glad expressing happy emotion

(d) Gesture sad expressing sad emotion

(e) Scared expressing scared emotion

(f) Gesture stiff expressing angry emotion

(g) Gesture surprise expressing surprise emotion

Figure 5: Few frames of PSLMNM dataset for each seven emotion-based gesture

4.2 Implementation Details

The system used for training and testing was the 8th generation Intel Core i7processor with 32 GB RAM and Nvidia GeForce GTX 1650 GPU. The model was implemented on Ubuntu with Python3 and TensorFlow. A leaky Rectified Linear Unit (leaky ReLU) was applied as an activation function in testing and training the model. The model was trained by keeping the initial learning rate of 0.0005,

momentum 0.93, optimizer weight decay 0.001, and batch size of 64 with Adam as an optimization function and Xavier method for initializing weight of the convolutional kernels. The cross-entropy loss function was used, which is stated in Eq. (4). Here 'a' is the actual value while 'b' is the predicted value by the model.

The loss was calculated as the difference in entropy between the expected and actual values. Fig. 6 graph between iteration along the x-axis and loss on the y-axis shows that the loss decreases gradually as the number of iterations increases.

$$H(a,b) = -\sum a(x)\,\mathrm{logb}\,(\mathrm{x})\,. \tag{4}$$

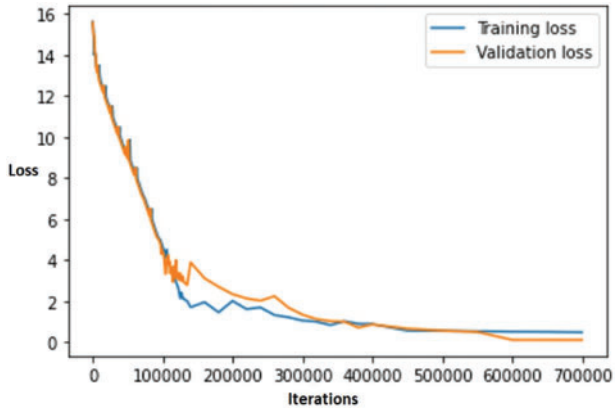


Figure 6: Error loss graph of training and validation

The proposed model achieved 86.12% training accuracy and 82.66% testing accuracy at 700000 iterations, as shown in Fig. 7. Our model can recognize the seven different actions formed by individuals. In Fig. 8, the top predicted activities for all categories using our model are depicted. It is noticeable that our proposed architecture can exploit the dominated expression to recognize the sign gesture.

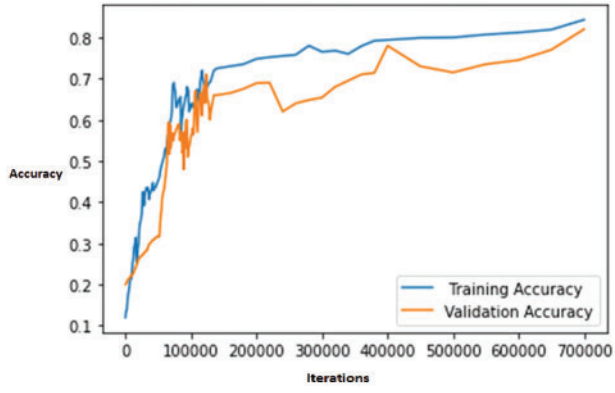


Figure 7: Training and validation accuracy graph

(a) Bad (disgust expression) (b) Best (neutral expression) (c) Glad (happy expression)

(d) Sad (sad expression) (e) Scared (scared expression) (f) Surprise (surprise expression)

(g) Stiff (angry expression))

Figure 8: Top predicted videos from the validation set using SLATN

It is already noticeable that our model can exploit the dominant facial expression to recognize any sign gesture. Due to this specific, our model predicts a few videos are misclassified. In Fig. 9. (a) Video from a surprise gesture is classified in Glad as in maximum frames; the facial expression represents a happy gesture which creates a failure mode. (b) Video from Bad gestures is misclassified in Best Gestures due to dominated facial expression. The hand's orientation is wrong and predictable in the key-frames. (c) Gestures belonging to the Glad class art are misclassified due to the wide opening of eyes, which dominates towards surprise as well as hand gesture is similar to surprise expression in most frames.

(a) Surprise Gesture (b) Bad Gesture (c) Glad Gesture

Figure 9: Misclassified videos from the validation set using SLATN

4.3 Result and Experiments

The proposed model is compared with the conventional recognition and classification convolutional neural network models. For testing these models, 25 samples from the validation set are passed to each model for prediction.

The Tab. 1 Shows the evaluation of models based on their performance, the computation required to process Giga FLOPS (GFLOPS), and parameters. The results show that the proposed SLATN model outperforms conventional models significantly with better performance and reducing the required computational power. It is important to notice that the proposed model had no frozen backbone architecture, unlike a few models used in the comparison, which eventually makes such models slower. On the other hand, at the intermediate nodes of the network, SLATN enables point-to-point fusion while training, allowing it to learn and perform better, outperforming other models.

Table 1: Different models comparison

Different model	mAP(%)	GFLOPs	Param (M)
I3D + super-events [27]	19.41	4446.15	26.18
ViVit [28]	18.55	3992.00	47.00
MViT (deep network) [29]	47.7	7080.00	53.00
I3D + super-events + TGM [30]	22.56	4446.75	28.28
ViT-B-VTN [31]	79.80	4218.00	114.00
I3D + STGCN [32]	19.09	4450.94	29.18
SLATN (Proposed model)	**66.10**	**94.13**	**6.80**

5 Conclusion

This paper proposes a novel, time-efficient action transformer network for sign language recognition named SLATN. SLATN is able to recognize and distinguish manual and non-manual gestures, where not only hand movements are considered, but the head and shoulder movements and expressions are localized in video sequences. Current architecture is expending a Transformer-style structural design as a "base network" to extract features from a spatiotemporal domain using our newly developed dataset PkSLMNM. Further, a "head network" simultaneously emphasizes hand movement and facial expression, using its attention mechanism to create tight bounding boxes around classified gestures. The network identifies gestures such as bad, best, sad, glad, scared, stiff, and surprised with good accuracy of 86.12% for training and 82.66% for testing and speed compared to the other present state-of-the-art algorithm providing mAP of 66.10% and Giga FLOPs as 94.13. Proposed system can be used in offices, hospitals, educational institutes, law enforcement, surveillance, etc. to provide a bridge between mute and normal persons. As the architecture has a limitation that sometimes the higher appearance features are missing in the recorded video leading to motion blur. In the future, the availability of resources will make it possible to improve the quality but apprehend delicate appearance and motion information using some memory constrictions environment. As well as, as a future direction we can expand the dataset, we can contribute to handle the blur or low feature frames in video frame selection and segmentation domains, we also has a future dimension to extract face as Region of Interest (ROI) and after extracting the features of facial gestures and simultaneously

from the other body or hand movement information, applying features fusion to achieve best accuracy as well as efficiency.

Contributions: Conceptualization, S.J., and S.R.; Methodology, S.J., and S.R.; Software, S.J.; Validation, S.J. and S.R.; Formal Analysis, S.J.; Investigation, S.J. and S.R.; Resources, S.J. and S.R.; Data Curation, S.J.; Writing—original draft preparation, S.J.; Writing—review and editing, S.R.; Visualization, S.J.; Supervision, S.R.

Acknowledgement: We acknowledge the contribution of organizations and people who participated in PkSLMNM ingenuity as participants, organizers, or evaluators.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. H. Organization, "World report on hearing," 2021. [Online]. Available: https://www.who.int/publications/i/item/world-report-on-hearing.
[2] M. Jebali, A. Dakhli and M. Jemni, "Vision-based continuous sign language recognition using multimodal sensor fusion," *Evolving Systems*, vol. 12, no. 4, pp. 1031–1044, 2021.
[3] A. Mehrabian, "Communication without words," *Communication Theory*, vol. 6, pp. 193–200, 2008.
[4] P. Kumar, P. P. Roy and D. P. Dogra, "Independent Bayesian classifier combination based sign language recognition using facial expression," *Information Sciences*, vol. 428, pp. 30–48, 2018.
[5] M. Deriche, S. O. Aliyu and M. Mohandes, "An intelligent arabic sign language recognition system using a pair of LMCs with GMM based classification," *IEEE Sensors Journal*, vol. 19, no. 18, pp. 8067–8078, 2019.
[6] R. Elakkiya, "Machine learning based sign language recognition: A review and its research frontier," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 7, pp. 7205–7224, 2021.
[7] R. Rastgoo, K. Kiani and S. Escalera, "Sign language recognition: A deep survey," *Expert System Applications*, vol. 164, pp. 113794, 2021.
[8] E. P. da Silva, P. D. P. Costa, K. M. O. Kumada and J. M. De Martino, "Facial action unit detection methodology with application in Brazilian sign language recognition," *Pattern Analysis and Applications*, vol. 24, pp. 1–17, 2021.
[9] M. T. Ubaid, T. Saba, H. U. Draz, A. Rehman and H. Kolivand, "Intelligent traffic signal automation based on computer vision techniques using deep learning," *IT Professionals*, vol. 24, no. 1, pp. 27–33, 2022.
[10] Y. Tao, S. Huo and W. Zhou, "Research on communication app for deaf and mute people based on facial emotion recognition technology," in *2020 IEEE 2nd Int. Conf. on Civil Aviation Safety and Information Technology (ICCASIT)*, Weihai, China, pp. 547–552, 2020.
[11] A. Hassouneh, A. M. Mutawa and M. Murugappan, "Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods," *Informatics in Medicine Unlocked*, vol. 20, pp. 100372, 2020.
[12] S. T. Ly, G. -S. Lee, S. -H. Kim and H. -J. Yang, "Emotion recognition via body gesture: Deep learning model coupled with keyframe selection," in *Proc. of the 2018 Int. Conf. on Machine Learning and Machine Intelligence*, Ha Noi, Viet Nam, pp. 27–31, 2018.
[13] C. M. A. Ilyas, R. Nunes, K. Nasrollahi, M. Rehm and T. B. Moeslund, "Deep emotion recognition through upper body movements and facial expression," in *16th Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications VISIGRAPP (5: VISAPP)*, Setúbal, Portugal, pp. 669–679, 2021.

[14] A. Kratimenos, G. Pavlakos and P. Maragos, "Independent sign language recognition with 3D body, hands, and face reconstruction," in *ICASSP 2021–2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Koya, Iraq, pp. 4270–4274, 2021.
[15] M. Jebali, P. Dalle and M. Jemni, "Sign language recognition system based on prediction in human-computer interaction," in *Int. Conf. on Human-Computer Interaction*, Heraklion, Crete, pp. 565–570, 2014.
[16] A. Boukdir, M. Benaddy, A. Ellahyani, O. E. Meslouhi and M. Kardouchi, "3D gesture segmentation for word-level arabic sign language using large-scale RGB video sequences and autoencoder convolutional networks," *Signal Image Video Processing*, vol. 16, pp. 1–8, 2022.
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hong Kong, China, pp. 1–9, 2015.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 20, pp. 5998–6008, 2017.
[19] N. Parmar, V. Ashish, U. Jakob, K. Lukasz, S. Noam *et al.* "Image Transformer." in *35th Int. Conf. on Machine Learning*, Stockholm, Sweden, pp. 4055–4064. PMLR, 2018.
[20] P. Ramachandran, B. Zoph and Q. V. Le, "Searching for activation functions," ArXiv Preprint ArXiv171005941, 2017.
[21] "Deaf reach schools and training centers in Pakistan", 2022. [Online]. Available: https://www.deafreach.com/.
[22] "PSL: Pakistan sign language", 2022. [Online]. Available: https://psl.org.pk/.
[23] A. Imran, A. Razzaq, I. A. Baig, A. Hussain, S. Shahid *et al.,* "Dataset of Pakistan sign language and automatic recognition of hand configuration of urdu alphabet through machine learning," *Data in Brief*, vol. 36, pp. 107021, 2021.
[24] "Pakistan sign language dataset-Open Data Pakistan", 2022. [Online]. Available: https://opendata.com.pk/dataset/pakistan-sign-language-dataset.
[25] H. Zahid, M. Rashid, S. Hussain, F. Azim, S. A. Syed *et al.,* "Recognition of urdu sign language: A systematic review of the machine learning classification," *PeerJ Computer Science*, vol. 8, pp. e883, 2022.
[26] S. Javaid, "PkSLMNM: Pakistan sign language manual and non-manual gestures dataset," 2022, [Online]. Available: https://data.mendeley.com/datasets/m3m9924p3v/1 (https://doi.org/10.17632/m3m9924p3v.1.).
[27] A. J. Piergiovanni and M. S. Ryoo, "Learning latent super-events to detect multiple activities in videos," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, pp. 5304–5313, 2018.
[28] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic *et al.,* "Vivit: A video vision transformer," in *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 6816–6826, 2021.
[29] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan *et al.,* "Multiscale vision transformers," in *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 6804–6815, 2021.
[30] A. J. Piergiovanni and M. Ryoo, "Temporal Gaussian mixture layer for videos," in *Int. Conf. on Machine Learning*, long beach, California, pp. 5152–5161, 2019.
[31] D. Neimark, O. Bar, M. Zohar and D. Asselmann, "Video transformer network," in *IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, pp. 3156–3165, 2021.
[32] P. Ghosh, Y. Yao, L. Davis and A. Divakaran, "Stacked spatio-temporal graph convolutional networks for action segmentation," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Snowmass Village, CO, USA, pp. 576–585, 2020.