Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666

Volume 10 Issue 40 (September 2025) PP. 152-169
DOI: 10.35631/JISTM.1040011

Journal of Information Systems and Technology Management (JISTM)
eISSN: 0128-1666

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT (JISTM)
www.jistm.com

GLOBAL ACADEMIC EXCELLENCE

AUGMENTED REALITY INTEGRATED ROBUST FUSION MODEL FOR SIGN LANGUAGE RECOGNITION USING COMPUTER VISON AND MACHINE LEARNING

A F M Saifuddin Saif [1*], Zainal Rasyid Mahayuddin[2*], Kevin Van Winkle[3]

[1] Department of Computing, Information and Mathematical Sciences, and Technology (CIMST), Chicago State University, USA
Email: asaif@csu.edu

[2] Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 UKM, Bangi, Selangor, Malaysia
Email: zainalr@ukm.edu.my

[3] Department of English & World Languages, Colorado State University Pueblo, Pueblo, CO 81001, USA
Email: kevin.vanwinkle@csupueblo.edu

[*] Corresponding Author

Article Info:

Article history:
Received date: 26.06.2025
Revised date: 10.07.2025
Accepted date: 06.08.2025
Published date: 14.09.2025

To cite this document:

Saif, A. F. M. S., Mahayuddin, Z. R., & Van Winkle, K. (2025). Augmented Reality Integrated Robust Fusion Model for Sign Language Recognition Using Computer Vison and Machine Learning. *Journal of Information System and Technology Management, 10* (40), 152-169.

DOI: 10.35631/JISTM.1040011

Abstract:

Sign language recognition (SLR) interprets sign language into text, bridging the communication gap between the deaf-mute community who use sign language and those who do not. Recent advancements in computer vision, deep learning, and augmented reality have shown significant progress in the field of motion and gesture recognition, however, large variations in hand actions, facial, and body postures, and the absence of region-specific datasets prevent universally accurate effective sign language recognition. This research developed an efficient model for SLR which includes an RGB-MHI attention module, and the Faster R-CNN deep learning architecture integrated with augmented reality. The proposed model was validated on two benchmark datasets, achieving an accuracy of 98.97% on the AUTSL dataset and 96.7% on the BosphorusSign22k dataset. Furthermore, the model was tested on a self-created dataset named "Amar Vasha" based on Bangla Sign Language (BdSL) to ensure cross-domain adaptability. Experimental results demonstrate that the proposed model achieves state-of-the-art performance on all three benchmarks.

Keywords:

Computer Vision, Machine Learning, Augmented Reality, Sign Language Recognition

Introduction

Human-computer interaction (HCI) applied to sign language recognition (SLR) can play a crucial role in enhancing communication between individuals who use sign language and those who communicate verbally. This form of machine interpretation relies on semantic cues encompassing various manual and non-manual visual modes of communication, such as hand gestures, body orientation and movement, facial expressions, eye gaze, and mouth shapes. Typically, SLR systems use red, blue, green-motion history images (RGB-MHI) attention modules fused with the Faster R-CNN deep learning architecture to create and enhance classification of these signs; however, advancements in computer vision and deep learning have significantly improved the processing of these diverse semantic features, enabling vision-based recognition to automatically interpret sign language with much more nuance, speed and accuracy. This, in turn, has enhanced the ability of sign language users to communicate with others who are not fluent in sign language.

Despite recent and rapid growth of the SLR tools for translating sign language into verbal language, the scope, number of modalities and variety of non-verbal, manual sign-based languages and their associated lexicons have impeded the creation of a consistent, universal SLR, particularly compared with conventional action recognition (Han et al., 2022). There are currently 300 different sign-based languages used by more than 70 million people around the world. In some of those languages, a facial expression is a common modality of communication, especially to convey emotions (Han et al., 2022). In others still, arms and hand gestures are commonly used to express meaning, while in others body motion is typical (Rastgoo et al., 2022). Furthermore, different signers using the same language may show a great deal of variance in the way they sign the language; speed, body motion, and whether the user is right or left-handed can all influence reception (Moryossef et al., 2021). The use of similar gestures can impose various meanings depending on the number of repetitions a user performs (Moryossef et al., 2021).

Further improvement of SLR systems using augmented reality requires increasing the size and scope of datasets used by the systems and to collect data from all the various sign languages used around the world (Mahayuddin and Saif, 2021). Doing so, however, is logistically complicated and cost prohibitive. A potential remedy to these limitations, however, is repurposing and validating existing datasets (De Coster et al., 2021). This study incorporates an innovative fusion model for sign language recognition using the previously collected AUTSL (Sincan and Keles, 2020; Boháček and Hrúz, 2022) and BosphorusSign22k (Özdemir et al., 2020) datasets, in combination with a dataset for Bangla sign language created for this study, named the Amar Vasha (AV) dataset.

This new fusion model incorporates an RGB-MHI attention module, setting itself apart from previous research by not relying on explicit modalities. The RGB-MHI attention module was designed to summarize entire videos in a single frame without the need to explicitly recognize specific parts, such as a user's hands or face. The proposed RGB-MHI attention module provides robust semantic cues for overall recognition, while reducing temporal loss through the preservation of motion history information for each frame in three different channels.

This research integrates the RGB-MHI attention module with the Faster R-CNN deep learning framework, creating a multi-model assembly to enhance accuracy in sign language recognition using augmented reality. The model achieved an accuracy of 98.5% for the AUTSL dataset and 95.6% for the BosphorusSign22k dataset, surpassing previous research in this area. Additionally, the self-created AV dataset, was developed to ensure its adaptability across various domains. An accuracy of 99.7% was achieved, surpassing other datasets for the same language.

The remainder of the paper is organized as follows: Section 2 situates the new model within related research. Section 3 expands upon the proposed research methodology. Section 4 provides comprehensive experimental results for robust validation of the proposed methodology. Finally, section 5 concludes with explicit discussion of results and offers some potential directions for future research.

Background Study

Early research in SLR systems used glove-based electrochemical devices to ascertain sign symbols. While useful, such systems altered sign language users' normal conveyance of signs, as the gloves were unnatural and cumbersome, making interpretation susceptible to misclassification (Karmokar et al., 2012). The development of various subsequent research methods, bolstered by digital video and virtual reality, has enabled more natural recognition of signs. These latter methods also allowed for the examination of additional communication channels, such as facial expressions and body posture. Figure 1 outlines the various methods researchers have used.

In addition to methods that allowed sign language users to communicate more naturally, researchers have sought to investigate the use of the supplementary channels of visual communication used in tandem with manual signing. For instance, Viegas et al. (2022) sought to ground grammatical and semantic functions of facial cues by modeling the relationship between text, gloss, and users' facial expressions. While this model improved the overall quality of automatically generated sign language, challenges remained. Finger movement was difficult to ascertain, making complex hand shapes harder to recognize. Additionally, tools capable of mitigating the loss of facial expression caused by occlusion were determined to be needed.

Similarly, the resemblance between words in sign language poses a challenge for translation algorithms. Abdullahi and Chamnongthai (2022) addressed this issue by employing fast fisher vector (FFV) to select and encode in deep learning bidirectional long short-term memory (Bi-LSTM). They utilized orientation angles and prosodic features to differentiate between similar sign patterns. Despite improvements in accuracy, the FFV-Bi-LSTM faced challenges in learning subtle hand motion differences for similar patterns, resulting in biases and misclassifications. Furthermore, the proposed method struggled to handle similar sign patterns, introducing a multi-feature problem.

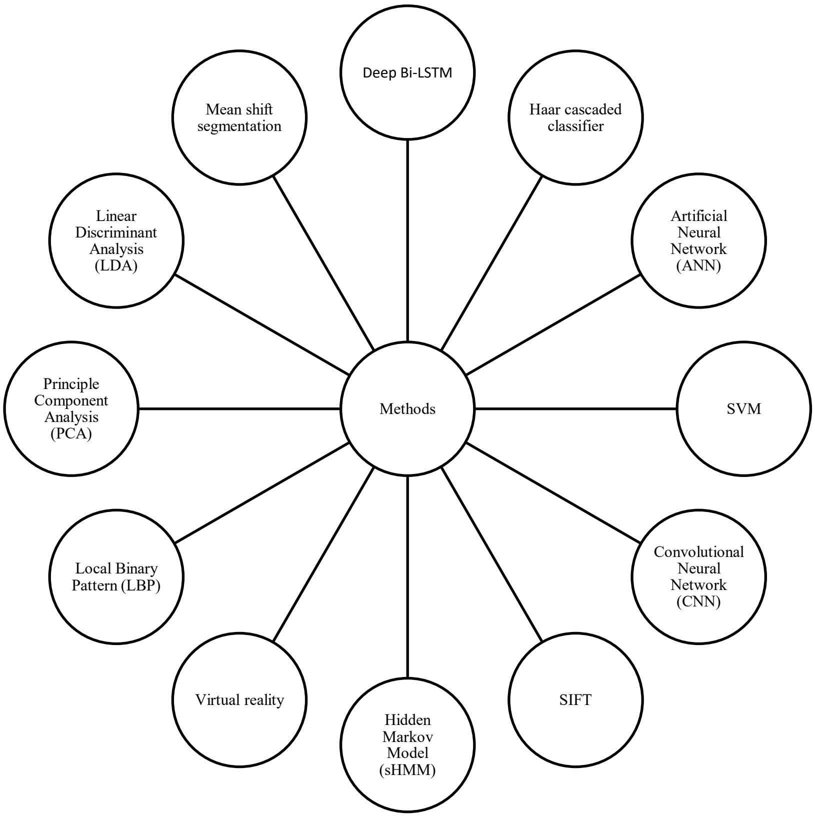
Deep Bi-LSTM
Haar cascaded classifier
Mean shift segmentation
Linear Discriminant Analysis (LDA)
Artificial Neural Network (ANN)
Principle Component Analysis (PCA)
Methods
SVM
Local Binary Pattern (LBP)
Convolutional Neural Network (CNN)
Virtual reality
SIFT
Hidden Markov Model (sHMM)

Figure 1: Existing Methods for Sign Language Recognition.

Jiang et al. (2021) proposed a skeleton aware multi-modal SLR. This method aimed to address challenges in understanding complex hand gestures, body posture, and facial expressions. The method introduced a unique skeleton graph using whole-body key points obtained from a pretrained pose estimator and relied on multiple depth modalities.

Barbhuiya et al. (2021) employed convolutional neural networks (CNNs) to identify static signs across different users. They utilized a customized pre-trained backbone network and a multiclass support vector machine (SVM) classifier, specifically targeting American Sign Language letters and numbers. The researchers achieved an accuracy of 99.82%, outperforming other methods. This cost-effective approach proved to enhance human-computer interaction on a regular computer; however, the researchers observed that omitting discriminative features resulted in a higher misclassification rate for similar sign patterns which was also observed by Mahayuddin and Saif (2021) during classification of data using augmented reality. Rastgoo et al. (2021) integrated multiple features for hand sign recognition, utilizing multimodalities, such as deep hand features and hand pose features. Their proposed method involved extracting features from various modalities. To enhance semantic information, the researchers introduced an RGB-MHI attention module, leveraging motion history images. The RGB-MHI module's format enriched features, addressing challenges like skin color similarity and body posture related to sign language recognition. Additionally, the proposed RGB-MHI attention module was fused with the Faster R-CNN deep learning framework to improve classification for signs at a far distance and those that require both hands. These enhancements were subsequently validated on three comprehensive datasets.

For the most part, the scope of research on SLR has been limited to American Sign Language (ASL) (Naglot and Kulkarni, 2016), French Sign Language (FSL) (Naglot and Kulkarni, 2016), Danish Sign Language (DSL) (Naglot and Kulkarni, 2016), British Sign Language (BSL) (D RAJ and Jasuja, 2018), and Indian Sign Language (ISL) (Ekbote and Joshi, 2017). Despite its approximately 3 million users, Bangla Sign Language (BdSL) has not been rigorously investigated for recognition, with only a much smaller subset of Sign Language Recognition (SLR) research conducted on it. For example, Karmokar et al. (2012) implemented a Neural Network Ensemble (NNE) with real-time input from a webcam to investigate BdSL. Their findings revealed that the NNE struggled in conditions of insufficient light, low-quality images, or when the user's skin color closely resembled that of the environment.

In a related study, Rahaman et al. (2014) utilized cascaded classifiers based on Haar-like features to recognize BdSL hand signs. Their approach involved detecting the hand area, converting signs into binary images using hue and saturation, and classifying them with a pre-trained K-Nearest Neighbor classifier (Mahayuddin and Saif, 2022). Although they achieved a notable accuracy of 98.17% for vowels, the accuracy for consonants was unsatisfactory. Furthermore, efforts to address skin color (Mahayuddin and Saif, 2021) and hand segmentation (Saif and Mahayuddin, 2020) did not yield satisfactory outcomes. Rahaman et al. (2015) later utilized Canny edge detection to encode Vector Contours (VC) for contour-based sign language recognition from gray images of BdSL sign words. They created a feature space for each sign word using the Auto-Correlation Function (ACF), despite the increased computation complexity. Meanwhile, Ahmed and Akhand (2016) employed an Artificial Neural Network (ANN) to determine the relative tip positions of a BdSL sign language user's fingers in two-dimensional space. While effective, this method struggled to accurately recognize signs at long distances.

In another approach, Hoque et al. (2018) adopted Faster-r CNN for real-time detection of BdSL recognition. While effective on their BdSLImset dataset, it proved limited when applied to other sign systems. Similarly, Sharmin et al. (2018) explored real-time usage of BdSL, employing color segmentation and converting frames into YCbCr. Despite using Hu moment invariants for feature extraction and SVM for classification, this method faced difficulties detecting the hand region when the background resembled skin color.

Addressing challenges related to users' skin color, Aziz et al. (2017) utilized mean shift segmentation and geometric hashing for one-handed signs but encountered issues in the presence of skin-colored objects. Yasir et al. (2015) used Scale Invariant Feature Transform (SIFT) for their developed Bag of Feature (BoF) for BdSL sign language recognition. Although, they used SVM classifier on vocabulary set, average accuracy for classification was not satisfactory enough comparing with existing research. Shanta et al. (2018) also applied a SIFT approach for BdSL sign language detection, using CNN for classification and achieving accuracy; however, they did not include two-handed gestures during validation and struggled with recognizing sign language from body posture. In contrast, Yasir and Khan (2014) incorporated two-hand gestures in BdSL sign language, applying principal component analysis (PCA) (Saif and Mahayuddin, 2022) and LDA to reduce dimensionality and transform images. While their method incurred additional computation costs due to a large framework, it achieved better accuracy. Yasir et al. (2017) used leap motion controller (LMC) based on virtual reality (Saif and Mahayuddin, 2021) to track continuous motion, achieving low error rates. Santa et al. (2017) employed Local Binary Pattern (LBP) for feature extraction and SVM for classification from video of BdSL sign language users. Despite converting RGB data into the YCbCr color space for skin color segmentation, their method faced challenges when dealing with multiple hands.

Proposed Method

The proposed research methodology is comprised of four components: data pre-processing, RGB-MHI attention module, Faster R- CNN, and fusion of RGB-MHI attention module with Faster R-CNN shown in Figure 2. The RGB-MHI attention module was formed from the feature map received from base network VGG-16 to extract strong semantic information from RGB frame and later fused with Faster-RCNN for final classification. Prior data pre-processing was conducted, followed by data augmentation for improved performance of the proposed model. Further details are illustrated in subsequent sections.

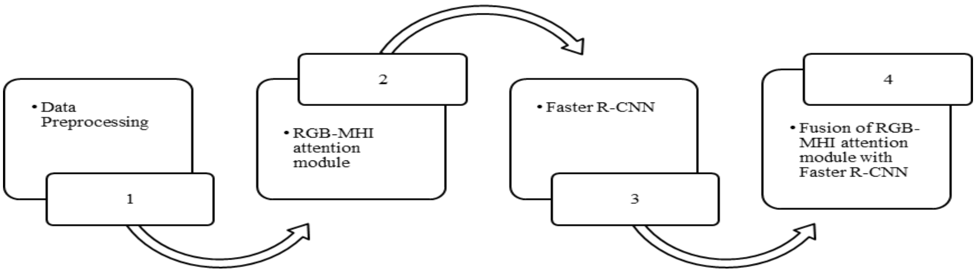
• Data Preprocessing
1
2
• RGB-MHI attention module
• Faster R-CNN
3
4
• Fusion of RGB-MHI attention module with Faster R-CNN

Figure 2: Components of Proposed Fusion Model.

Data Preprocessing

A VGG-16 backbone network was used for sufficient semantic information extraction from RGB data and to provide enough information for the RGH-MHI attention module in the subsequent step. In certain RGB data, the presence of another person moving behind the signer prompted the selection and prioritization of the person closest to the camera, specifically focusing on the one with the relatively largest bounding box. Consequently, expansion of the bounding box on both left and right sides was accomplished and all the RGB data were cropped into squares. RGB-MHI data were generated for every video frame and the images were cropped into a square format using bounding box information obtained from a pretrained Faster R-CNN model applied to the respective datasets. For all videos, 33 frames were fixed based on uniform sampling.

RGB-MHI Attention Module

Attention mechanisms, implemented within deep learning architectures, have shown promise at enhancing the classification performance in computer vision tasks. In sign language recognition, attention mechanisms can serve as a viable alternative to segment either spatial or temporal parts of the stream. This approach helps in constructing robust models by focusing on specific cues during the recognition process. This research utilized an RGB-MHI attention module with Faster R-CNN deep learning architecture to focus on semantic cues extracted in the previous step. The proposed model used only RGB data without any explicit part for recognition, such as the user's hands or face. RGB-MHI was formed to summarize the entire video in a single frame to represent compact and relevant semantic pattern and to feed into the Faster R-CNN deep learning framework.

MHI is gray scale image format to capture brighter moving pixels (Bobick and Davis, 2001). Proposed RGB-MHI includes magnitude and phase information inspired by previous research (dos Santos et al., 2020). Each frame is split into three equal parts for calculating motion history separately. In this context, the R-channel preserves motion history data from the initial temporal region; the G-channel preserves from central one; and the B-channel preserves from the last temporal region. Thus, preservation of motion information into three separate channels helped temporal information loss. MHI images were calculated by adding the absolute differences in pixel intensities between consecutive frames, as described in Eq. (1). Here S denotes the total number of frames in video, f_t denotes t video frame, coordinates of pixels are denoted by (a; b), absolute difference of the weight is represented by K calculated as K = t/S.

$$M(a,b) = \sum_{t=2}^{N} f_{t-1}(a,b) - f_t(a,b)|\mathrm{K} \qquad (1)$$

The proposed RGB-MHI attention module used pretrained VGG-16 on the ImageNet (dos Santos et al., 2020) dataset and was then fine-tuned for three sign language datasets. The VGG-16 included a fully connected (FC) layer consisting of 1024 neurons and incorporated a ReLU non-linearity. Subsequently, to mitigate the risk of losing details in the spatial context effects for the MHI images, a final FC layer, utilizing a SoftMax function, followed. Optimal results were attained when all the convolutional layers in the pretrained VGG-16 model were employed.

Faster R-CNN based Flatten Feature Maps Using Semantic Feature Extraction

For strong semantic cues, such as texture, edge and color from an RGB image, Faster R-CNN deep learning architecture was used to locate regions of interest (ROI). The use of the region proposal network (RPN) in Faster R-CNN results in the selection of fewer windows with practical scales and aspect ratios, influencing the identification of positively classified objects. Accordingly, a non-maximal suppression technique is applied for further refinement. Proposal costs were minimized by employing the RPN, which shares convolutional layers. Subsequently, an additional convolutional layer was utilized to regress region bounds and objectness scores based on the foreground position.

The RPN acted as an attention detector that was fused with the RGB-MHI attention module in order to determine optimum bounding boxes across scales and aspect ratios in the prediction layer. Among convolutional layers, 13 layers + 13 layers +4 layers were used to extract feature maps for sharing with subsequent RPN layers and fully connected layers. The RPN used convolutional layers to generate foreground anchors based on the SoftMax classifier and bounding box regression offsets for proposal calculation. Max pooling was used to produce fixed size, 7*7 feature maps aligned with window size. If the longer side of the image was excessively large, it was constrained to a specified size, preserving the aspect ratio. Faster R-CNN extracted features by projecting regions through the RPN onto the deep convolutional layer, specifically conv-5. This layer, from the VGG-16 architecture, is downsampled by a factor of 32 compared to the original input size. This is done to obtain appropriate features for the prediction layer. The overall Faster R-CNN architecture employed by the proposed model is illustrated in Figure 3.

Fusion of Faster R-CNN and RGB-MHI Attention Module

The RGB-MHI attention module was fused with Faster R-CNN deep learning framework as multi-model assembles to achieve higher accuracy. Two weights, l_1 and l_2, were assigned in the last layers before SoftMax to calculate weighted sum for final prediction as discussed in Eq. (2) and shown in Figure 3.

$$\text{Prediction} = \text{softmax}\ (l_1 x_1 + l_2 x_2) \qquad (2)$$

Balance was achieved using Single Shot Detector (SSD) (Mahayuddin and Saif, 2019). To compute the feature map, SSD operates on the output of the classification layer from Faster R-CNN only once per given frame. SSD applied a small 4x4 convolutional kernel to the feature map to predict bounding boxes and their corresponding probability estimates. Anchor boxes were also used at various aspect ratios in order to learn the offset instead of learning the box. Each convolutional layer operated at a different scale, so objects with various scales could be detected with higher prediction rate. In this context, Unity3D environment was used for recognition object in augmented reality which is a cross platform. After that, ARCore solution was used to integrate augmented reality for mobile phone in order to make the recognition more convenient and to employ the sensor fusion to track planar surfaces. Besides, data augmentation was applied in the RGB-MHI formation in both the training and the validation sets. The bounding box from either the left or right was expanded to prevent the signer's appearance in the center of the videos. This required vertical shifting of the bounding box randomly within limited pixel constraints. Afterwards, horizontal flip was applied to accommodate the proposed model for both hands.

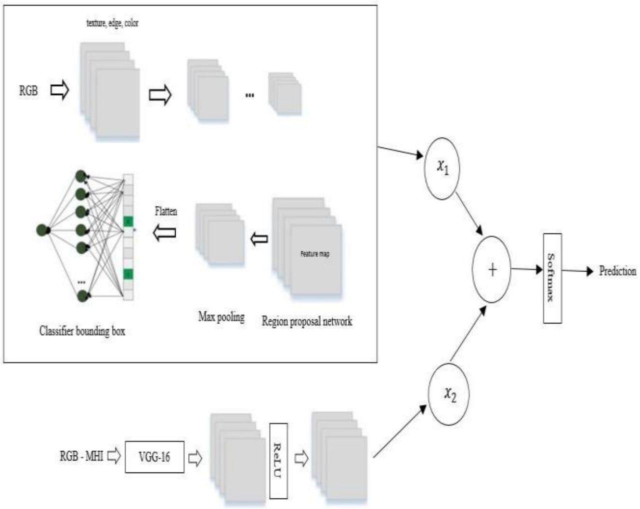

Figure 3: Proposed Fusion Model for Sign Language Recognition.

Experimental Results and Discussion

Experimentation of the proposed fusion model was conducted on an average computer system: an Intel® Core™ i7-6100U processor with 16 GB RAM running Windows 11.

Datasets

Three datasets are used for validation: AUTSL (Sincan and Keles, 2020; Boháček and Hrúz, 2022), BosphorusSign22k (Özdemir et al., 2020), and the Bangla sign language dataset developed for this study and named the Amar Vasha (AV) dataset. The AUTSL dataset, comprised of 226 signs and 36,302 video samples, is characterized as a large-scale and independent sign system, as opposed to being an isolated dataset. Among the 43 sign users in this dataset, 31 are training set. The validation set was six, and the remaining six were used in the test set. This dataset incorporated 20 distinct backgrounds, serving as a challenge to guarantee the signer-independent evaluation of the proposed method in this research. Various backgrounds that included moving trees and pedestrians behind the sign user were used.

The BosphorusSign22k dataset (Özdemir et al., 2020) is comprised of 744 signs with 22,542 video samples depicting health and finance scenarios, as well as other common activities where a sign language user would need to communicate with a non-sign language user. Six signers contributed to this dataset, with one reserved for testing. One signer was allocated to the validation set for this particular dataset.

Six Bangla signs were used in the AV dataset. Three of them were alphabetic signs (অ, আ, ও) and three were numeric signs (১, ২, ৩). More than a thousand images captured in 640x480 pixels of each sign were used to create the AV dataset. After normalization, 700 images with 224x224 pixels were used for each sign. In total, 4000 images were used for training and 520 images were used for testing.

In all datasets, when videos contained multiple individuals, and pedestrians were passing by the signer, the algorithm selected the largest bounding box associated with a person. The number of frames was fixed at 32; however, some frames were skipped during sampling for the isolated videos. After cropping, a square RGB-MHI image was generated to obtain the bounding box information from Faster R-CNN. Data augmentation was applied to improve the variability in the training and validation sets. To avoid having sign users consistently appear in the center of the videos, bounding boxes were adjusted either to the left or right. To accomplish this, bounding boxes were vertically shifted for the signer to be in various vertical positions. In order to balance the positioning equally, horizontal flip was also applied to accommodate the sign user when they used both hands. All frames were resized to 224x224 for validation.

Training Details

This research used PyTorch library (Paszke et al., 2019) for implementing the proposed method. Optimum hyper parameters were determined based on experimentation, i.e. learning rate. An Adam optimizer with an initial learning rate of 1e-4 was also used. The learning rate was reduced to a 0.2 factor, in case no improvement was observed in the validation set until three epochs. The overall process was repeated several times until improvement could be observed. Each dataset was partitioned into training and testing data. For experimentation purposes, 90% of the images were assigned to the training set, while the remaining 10% were allocated to the testing set. The complete training process was observed until total loss decreased to 0.05. When total loss decreased to 0.05, training was stopped to avoid overfitting. If total loss increased rather than decreased, the training process was immediately stopped. When this stoppage occurred, the last step number was saved for later usage in the inference graph and to monitor the whole training process. The proposed model was trained on GTX 1080 GPU, where inference time for one image was approximately 2s with 2k proposals.

Experimental Results on AUTSL Dataset

To observe empirical performance with different depths and to validate the proposed model, the VGG-16 was implicated on a layer-by-layer basis. The best result was achieved when all layers of VGG-16 were used. Compared to the baseline method, which is presented with the AUTSL dataset (Sincan and Keles, 2020), the proposed method achieved 95.3% classification accuracy using single image RGB-MHI. This is higher than 2D-CNN with an attention based bidirectional LSTM model taking entire video frames. Also, since it only needs single pattern interpretation from a single image, using RGB-MHI image training was very fast. This was especially true when compared to 2D models, where entire video frames are used as inputs. Moreover, since there are more than 200 different sign classes in the AUTSL dataset, the 95.3% accuracy rate showed that a single RGB-MHI image was capable of learning important discriminative features to represent isolated signs. In order to increase the generalization capability of the proposed method, training was done with the augmented data. During the experiment, the classification accuracy increased from 95.3% to 97.7% when seven times more training data was generated through the use of data augmentation. In previous research using the AUTSL dataset (Sincan and Keles, 2020), participants were allowed to use validation labels to fine tune the complete methodology. The initial learning rate was set to 2e - 3, instead of 1e - 2. Since there was no validation set in this experiment, fine tuning was stopped when the training loss was reduced to the same level as the previous run. The proposed method, with the union of training and validation sets, obtained a 98.97% classification accuracy. Experimental results on AUTSL dataset are shown in Table 1.

Table 1: Experimental Results on AUTSL Dataset.

Proposed Method	Frame Size	Accuracy	Computational Time
Without augmentation	224x224	95.3%	0.11s
With augmentation	224x224	97.7%	0.13s
After fine-tuning	224x224	98.97%	0.12s

Table 2 shows existing research results on the AUTSL dataset compared against the proposed method. Various modalities were used in previous research to increase classification accuracy, i.e. skeleton joints, optical flow, and hand crops. The proposed method received better results using the same AUTSL dataset. Thus, the utilization of both the RGB image and the RGB-MHI images derived from the RGB image distinctly demonstrated the effectiveness of the RGB-MHI attention module for sign language classification, along with the benefits of data augmentation.

Table 2: Comparison With Existing Research Results on AUTSL Dataset.

Method	Modalities	Accuracy
MS-G3D (Vázquez-Enríquez et al., 2021)	RGB, skeleton joints, bones, joint motion, bone motion	96.15%
I3D-VLE-transformer (Gruber et al., 2021)	RGB, hand crops, pose, location vectors	95.46%
VTN-PF (De Coster et al., 2021)	RGB, hand crops, location vectors, skeleton	92.92%
OpenPose+Holistic (Moryossef et al., 2021)	Skeleton	81.93%
2DCNN+Attentioned BLSTM (Sincan and Keles, 2020)	RGB	49.22%
Proposed Method	RGB, RGB-MHI image	98.97%

Experimental Results on BOSPHORUSSIGN22K Dataset

The evaluation of the proposed method on the BosphorusSign22k (Özdemir et al., 2020) dataset yielded a classification accuracy of 78.3%. Drawing inspiration from prior research (Sincan et al., 2021; Vázquez-Enríquez et al., 2021), the parameters trained on the AUTSL dataset were used to examine the outcomes of transfer learning. The decision to use AUTSL pretrained data was motivated by two factors: first, it contains similar signs, albeit in different contexts with different signers; and second, as demonstrated in previous research (Vázquez-Enríquez et al., 2021), pretraining with AUTSL has proven effective for other sign languages, such as Spanish Sign Language.

The proposed model was fine-tuned using the union of the training and validation datasets, resulting in an accuracy rate of 93.6%. In light of the limited number of signers in the training data (four signers) and to enhance the generalization of the proposed model, an additional sign language user was included in the validation set. The utilization of pretrained datasets, specifically AUTSL and ImageNet, contributed to an increased classification accuracy of 96.7%. The experimental results using the BosphorusSign22k dataset are presented in Table 3.

Table 3: Experimental Results on BOSPHORUSSIGN22K Dataset.

Proposed Method	Frame Size	Pretrained datasets	Accuracy	Computational Time
RGB-MHI	224x224	ImageNet	78.3%	0.11s
Fine tune with train + validation data	224x224	AUTSL	93.6%	0.12s
Fine tune with train + validation data	224x224	AUTSL, ImageNet	96.7%	0.14s

The performance of the proposed model on the BosphorusSign22k dataset, as compared to results from previous research, is presented in Table 4. The achieved accuracy of 96.7% surpasses the performance of other existing methods on the BosphorusSign22k dataset. Notably, the proposed method did not rely on concise part extraction for segmentation as a modality. Instead, RGB-MHI was employed to assist in focusing on relevant parts for sign classification.

Table 4: Comparison With Existing Research Results on BOSPHORUS22K Dataset

Method	Modalities	Accuracy
MC3-18 Spatio-Temporal Sampling (Gökçe et al., 2020)	RGB, cropped hands, cropped face	94.94%
IDT (Özdemir et al., 2020)	HOG, HOF, MBH	88.53 %
MC3-18 Spatio-Temporal Sampling (Gökçe et al., 2020)	RGB	86.91%
MC3-18 (Özdemir et al., 2020)	RGB	78.85 %
Proposed Method	RGB, RGB-MHI image	96.7%

Experimental Results on Amar Vasha (AV) Dataset

The AV dataset used in this experimentation comprises both alphabetic and numeric signs, totaling 38 signs from the 51 Bangla alphabets (Islam et al., 2018). This research specifically focused on six signs—অ, আ, ও, ১, ২, and ৩—selected based on criteria such as the use of both hands, the placement of hands with respect to the body, posture, and facial expression. Among these signs, single-handed gestures represent ১, ২, and ৩, while double-handed gestures represent অ, আ, and ও. The choice of these six signs was made to test both single and double-handed signs.

The classification results for these signs exhibited fluctuations due to variations in the experimental environment and the distance of the background from foreground objects. Similar to the AUTSL and BosphorusSign22k datasets, a frame size of 224x224 was employed for the AV dataset, owing to its good texture and smoothness. The proposed method achieved recognition accuracy of 99.9% for অ, 99.8% for আ, and 98.4% for ও, as shown in Table 5. The lower accuracy for ও compared to অ and আ was attributed to its similarity with the sign আ.

Regarding numeric signs, the proposed model achieved recognition accuracy of 99.9% for ১, 98.2% for ২, and 99.7% for ৩. However, due to the similar angle of hand positions during training, the accuracy for ২ was lower than that for ১ and ৩. Notably, accuracy increased when the angle of the signer's hand position approached that during training. Additionally, due to the similarity of these signs with others in terms of shape, the accuracy for ও and ৩ classifications fluctuated. Experimental results are detailed in Table 5, and selected real-time video frame results are presented in Figure 4.

Table 5: Summary of Experimental Results on AV Dataset.

Object	**Frame Size**	**Accuracy**	**Computational Time**
অ	224x224	99.9%	0.12s
আ	224x224	99.8%	0.12s
ও	224x224	98.4%	0.13s
১	224x224	99.9%	0.12s
২	224x224	98.2%	0.12s
৩	224x224	98.8%	0.12s

The proposed method was compared against existing Bangla sign language recognition methods, as summarized in Table 6. In one study (Karmokar et al., 2012), 47 signs with 235 samples achieved a recognition rate of 93%. However, the neural network ensemble (NNE) exhibited reduced accuracy in challenging conditions. Another investigation (Mahayuddin and Saif, 2021), trained on 3600 samples for 36 hand signs revealed challenges particularly for consonants, attained a recognition rate of 96.46%.

Two artificial neural network (ANN) architectures (Ahmed and Akhand, 2016) achieved a recognition rate of 98.99% using a fingertip position-based approach. In contrast, Faster R-CNN (Hoque et al., 2018) achieved a slightly lower accuracy of 98.20%, attributed to limitations in handling similar sign patterns. Another study, employing a dataset of 34 signs with 1020 sample images (Sharmin et al., 2018), demonstrated a training accuracy of 98.24% and a testing accuracy of 87.79%; however, challenges were encountered in hand region detection against a skin-colored background. Several other studies utilized various datasets and methods, achieving accuracies ranging from 70% to 99% (Shanta et al., 2018; Santa et al., 2017; Islam et al., 2018; Hasan et al., 2017; Uddin and Chowdhury, 2016; Hasan et al., 2016; Kishore et al., 2015). It is noteworthy that the proposed research, employing 4200 samples for 6 signs, surpassed these state-of-the-art methods with an outstanding accuracy of 99.1%.

Table 6: Comparison With Existing Research Results.

Method Name	Recognition Rate (Percent)	Total Signs	Total Samples	Frame Rate
Neural Network Ensemble (NNE) (Boháček and Hrúz, 2022)	93	47	235	N/A
K-Nearest Neighbor (KNN) (Özdemir et al., 2020)	96.46	36	3600	93.55792 milliseconds
Artificial Neural Network (ANN) (Karmokar et al., 2012; Mahayuddin and Saif,2021; Ahmed and Akhand, 2016)	98.99 (Karmokar et al., 2012)	37	518	N/A
	96.5 (Mahayuddin and Saif,2021)	20	400	N/A
	95.1 (Ahmed and Akhand, 2016)	36	288	N/A
Support Vector Machine (SVM) (Sincan and Keles, 2020; Rahaman et al. (2014); Saif and Mahayuddin, 2020;Rahaman et al., 2015)	97.7 (Saif and Mahayuddin, 2020)	N/A	4800	N/A
	86.53 (Rahaman et al., 2015)	16	320	N/A
	94.26 for words and 94.49 for sentences (Rahaman et al., 2014)	25	4800	25 fps
	87.79 (Sincan and Keles, 2020)	34	1020	
Convolutional Neural Network (CNN) (Rastgoo et al., 2022; Mahayuddin and Saif, 2022)	90.63 (Rastgoo et al., 2022)	38	1700	N/A
	92 (Mahayuddin and Saif, 2022)	36	1800	N/A
Faster Region-based CNN (Faster R-CNN) (Islam et al., 2018)	98.2	N/A	N/A	90.03 milliseconds
Proposed Method	99.7%	6	4200	8

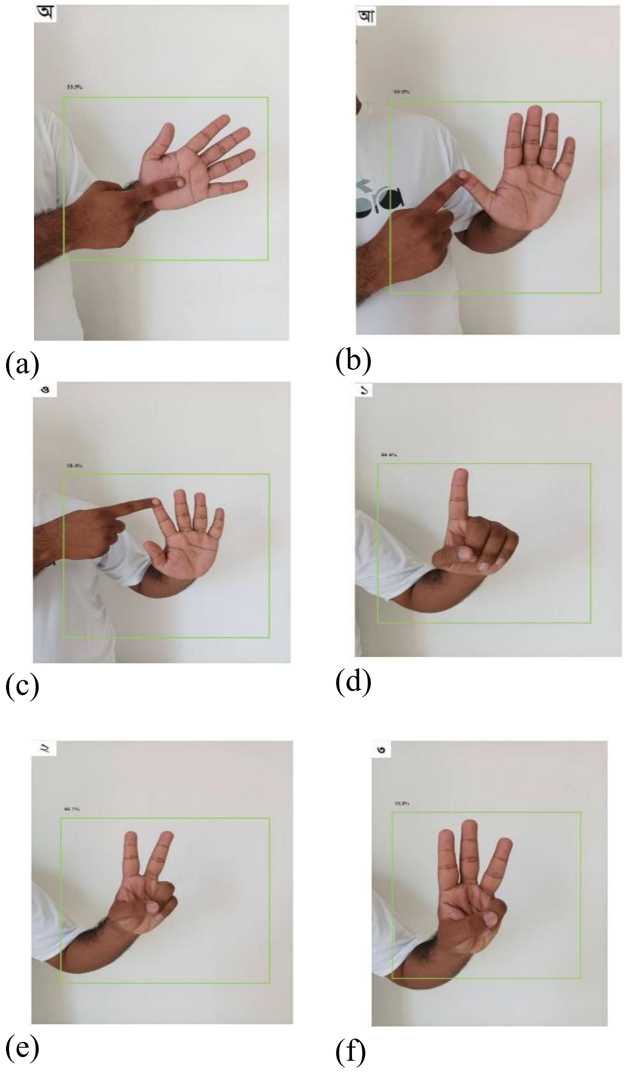

Figure 4: Bangla Sign Language Classification. (a) অ, with Accuracy 99%. (b) আ, with Accuracy 99.7% (c) ও, with Accuracy 99.8%. (d) আ, with Accuracy 99.7% (e) ২, with Accuracy 97.7% (f) ৩, with Accuracy 99.7%.

Conclusion

In this research, the entire video content was condensed into a single frame using the RGB-MHI attention module, without explicit focus on specific parts like hands or face. Three distinct channels in the form of motion history images were employed to capture robust semantic cues for comprehensive classification. The fusion of the RGB-MHI image with efficient machine learning framework named Faster R-CNN was performed to assemble a multi-model approach, resulting in higher accuracy. In this context, cross platform Unity3D environment was used to embed with the fusion model. Promising experimental outcomes were achieved on two publicly available datasets, AUTSL and BosphorusSign22k, with an accuracy of 98.97% and 96.7%, respectively, surpassing existing research benchmarks. Additionally, the proposed model underwent validation on a self-created dataset named Amar Vasha for Bengali Language, ensuring cross-domain adaptability. A noteworthy accuracy of 99.7% was attained for the Amar Vasha dataset, outperforming other datasets in the same language. The deaf and hearing-impaired community faces social barriers due to the limited adoption of sign language among the general population. To address this issue, a solution is proposed in the form of vision-based continuous sign language recognition. This technology aims to translate sign language gestures, thereby alleviating communication challenges. The future development of this model involves validation using diverse datasets that represent different sign languages and incorporate various deep learning architectures. The envisioned vision-based fusion model for sign language recognition is expected to play a crucial role in enhancing communication between the deaf and hearing-impaired community and the general population in their daily activities.

Acknowledgement

The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the "FTM1 - Peruntukan Dana Fakulti Teknologi dan Sains Maklumat, UKM". Besides, the author would like to thank the Department of Computing, Information and Mathematical Sciences, and Technology (CIMST), Chicago State University for various support for this research.

References

Abdullahi, S. B., & Chamnongthai, K. (2022). American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach. *IEEE Access, 10*, 15911-15923.

Ahmed, S. T., & Akhand, M. (2016). *Bangladeshi sign language recognition using fingertip position.* Paper presented at the 2016 International conference on medical engineering, health informatics and technology (MediTec).

Aziz, K. E., Wadud, A., Sultana, S., Hussain, M. A., & Bhuiyan, A. (2017). *Bengali Sign Language Recognition using dynamic skin calibration and geometric hashing.* Paper presented at the 2017 6th international conference on informatics, electronics and vision & 2017 7th international symposium in computational medical and health technology (ICIEV-ISCMHT).

Barbhuiya, A. A., Karsh, R. K., & Jain, R. (2021). CNN based feature extraction and classification for sign language. *Multimedia Tools and Applications, 80*(2), 3051-3069.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence, 23*(3), 257-267.

Boháček, M., & Hrúz, M. (2022). *Sign pose-based transformer for word-level sign language recognition.* Paper presented at the Proceedings of the IEEE/CVF winter conference on applications of computer vision.

D RAJ, R., & Jasuja, A. (2018). *British sign language recognition using HOG.* Paper presented at the 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS).

De Coster, M., Van Herreweghe, M., & Dambre, J. (2021). *Isolated sign recognition from rgb video using pose flow and self-attention.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

dos Santos, C. C., Samatelo, J. L. A., & Vassallo, R. F. (2020). Dynamic gesture recognition by using CNNs and star RGB: A temporal information condensation. *Neurocomputing, 400*, 238-254.

Ekbote, J., & Joshi, M. (2017). *Indian sign language recognition using ANN and SVM classifiers.* Paper presented at the 2017 International conference on innovations in information, embedded and communication systems (ICIIECS).

Gökçe, Ç., Özdemir, O., Kındıroğlu, A. A., & Akarun, L. (2020). *Score-level multi cue fusion for sign language recognition.* Paper presented at the Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16.

Gruber, I., Krnoul, Z., Hrúz, M., Kanis, J., & Bohacek, M. (2021). *Mutual support of data modalities in the task of sign language recognition.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Han, X., Lu, F., Yin, J., Tian, G., & Liu, J. (2022). Sign language recognition based on R (2+ 1) D With spatial–temporal–channel attention. *IEEE Transactions on Human-Machine Systems, 52*(4), 687-698.

Hasan, M., Sajib, T. H., & Dey, M. (2016). *A machine learning based approach for the detection and recognition of Bangla sign language.* Paper presented at the 2016 international conference on medical engineering, health informatics and technology (MediTec).

Hasan, M. M., Khaliluzzaman, M., Himel, S. A., & Chowdhury, R. T. (2017). *Hand sign language recognition for Bangla alphabet based on Freeman Chain Code and ANN.* Paper presented at the 2017 4th International Conference on Advances in Electrical Engineering (ICAEE).

Hoque, O. B., Jubair, M. I., Islam, M. S., Akash, A.-F., & Paulson, A. S. (2018). *Real time bangladeshi sign language detection using faster r-cnn.* Paper presented at the 2018 international conference on innovation in engineering and technology (ICIET).

Islam, M. S., Mousumi, S. S. S., Jessan, N. A., Rabby, A. S. A., & Hossain, S. A. (2018). *Ishara-lipi: The first complete multipurposeopen access dataset of isolated characters for bangla sign language.* Paper presented at the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP).

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). *Skeleton aware multi-modal sign language recognition.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Karmokar, B. C., Alam, K. M. R., & Siddiquee, M. K. (2012). Bangladeshi sign language recognition employing neural network ensemble. *International journal of computer applications, 58*(16), 43-46.

Kishore, P., Prasad, M. V., Prasad, C. R., & Rahul, R. (2015). *4-Camera model for sign language recognition using elliptical fourier descriptors and ANN.* Paper presented at the 2015 international conference on signal processing and communication engineering systems.

Mahayuddin, Z. R., & Saif, A. (2021). Vision based 3D Gesture Tracking using Augmented Reality and Virtual Reality for Improved Learning Applications. *International Journal of Advanced Computer Science and Applications*.

Mahayuddin, Z. R., & Saif, A. (2022). Moving Object Detection Using Semantic Convolutional Features. *J. Inf. Syst. Technol. Manag, 7*, 24-41.

Mahayuddin, Z. R., & Saif, A. S. (2019). *A Comprehensive Review Towards Appropriate Feature Selection for Moving Object Detection Using Aerial Images.* Paper presented at the Advances in Visual Informatics: 6th International Visual Informatics Conference, IVIC 2019, Bangi, Malaysia, November 19–21, 2019, Proceedings 6.

Moryossef, A., Tsochantaridis, I., Dinn, J., Camgoz, N. C., Bowden, R., Jiang, T., . . . Ebling, S. (2021). *Evaluating the immediate applicability of pose estimation for sign language* *recognition.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Naglot, D., & Kulkarni, M. (2016). *Real time sign language recognition using the leap motion controller.* Paper presented at the 2016 international conference on inventive computation technologies (ICICT).

Özdemir, O., Kındıroğlu, A. A., Camgöz, N. C., & Akarun, L. (2020). Bosphorussign22k sign language recognition dataset. *arXiv preprint arXiv:2004.01283*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems, 32*.

Rahaman, M. A., Jasim, M., Ali, M. H., & Hasanuzzaman, M. (2014). *Real-time computer vision-based Bengali sign language recognition.* Paper presented at the 2014 17th international conference on computer and information technology (ICCIT).

Rahaman, M. A., Jasim, M., Ali, M. H., & Hasanuzzaman, M. (2015). *Computer vision based bengali sign words recognition using contour analysis.* Paper presented at the 2015 18th International Conference on Computer and Information Technology (ICCIT).

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Hand pose aware multimodal isolated sign language recognition. *Multimedia Tools and Applications, 80*, 127-163.

Rastgoo, R., Kiani, K., & Escalera, S. (2022). Real-time isolated hand sign language recognition using deep networks and SVD. *Journal of Ambient Intelligence and Humanized Computing, 13*(1), 591-611.

Saif, A. S., & Mahayuddin, Z. R. (2020). Moment features based violence action detection using optical flow. *International Journal of Advanced Computer Science and Applications, 11*(11).

Saif, A. S., & Mahayuddin, Z. R. (2022). Vision based 3D Object Detection using Deep Learning: Methods with Challenges and Applications towards Future Directions. *International Journal of Advanced Computer Science and Applications, 13*(11).

Saif, A. S., Mahayuddin, Z. R., & Shapi'i, A. (2021). Augmented reality based adaptive and collaborative learning methods for improved primary education towards fourth industrial revolution (IR 4.0). *International Journal of Advanced Computer Science and Applications, 12*(6).

Santa, U., Tazreen, F., & Chowdhury, S. A. (2017). *Bangladeshi hand sign language recognition from video.* Paper presented at the 2017 20th International Conference of Computer and Information Technology (ICCIT).

Shanta, S. S., Anwar, S. T., & Kabir, M. R. (2018). *Bangla sign language detection using sift and cnn.* Paper presented at the 2018 9th international conference on computing, communication and networking technologies (ICCCNT).

Sharmin, S., Sultana, P., & Khan, M. I. (2018). Real time hand gesture recognition for Bangla character using SVM classifier. *Int. J. Comput. Appl., 180*(18), 24-28.

Sincan, O. M., Junior, J., Jacques, C., Escalera, S., & Keles, H. Y. (2021). *Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Sincan, O. M., & Keles, H. Y. (2020). Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access, 8*, 181340-181355.

Uddin, M. A., & Chowdhury, S. A. (2016). *Hand sign language recognition for bangla alphabet using support vector machine.* Paper presented at the 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET).

Vázquez-Enríquez, M., Alba-Castro, J. L., Docío-Fernández, L., & Rodríguez-Banga, E. (2021). *Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Viegas, C., Inan, M., Quandt, L., & Alikhani, M. (2022). Including facial expressions in contextual embeddings for sign language generation. *arXiv preprint arXiv:2202.05383*.

Yasir, F., Prasad, P., Alsadoon, A., Elchouemi, A., & Sreedharan, S. (2017). *Bangla Sign Language recognition using convolutional neural network.* Paper presented at the 2017 international conference on intelligent computing, instrumentation and control technologies (ICICICT).

Yasir, F., Prasad, P. C., Alsadoon, A., & Elchouemi, A. (2015). *Sift based approach on bangla sign language recognition.* Paper presented at the 2015 IEEE 8th international workshop on computational intelligence and applications (IWCIA).

Yasir, R., & Khan, R. A. (2014). *Two-handed hand gesture recognition for Bangla sign language using LDA and ANN.* Paper presented at the The 8th international conference on software, knowledge, information management and applications (SKIMA 2014).