AIMS Mathematics, 9(1): 2105–2122.
DOI:10.3934/math.2024105
Received: 22 October 2023
Revised: 06 November 2023
Accepted: 13 November 2023
Published: 20 December 2023

http://www.aimspress.com/journal/Math

Research article

Deep learning-based sign language recognition system using both manual and non-manual components fusion

Maher Jebali*, Abdesselem Dakhli and Wided Bakari

Computer Science Departement, University of Ha'il, P.O. Box 2440, Hail City, 100190, Saudi Arabia

* **Correspondence:** Email: maher.jbeli@gmail.com.

Abstract: Sign language is regularly adopted by speech-impaired or deaf individuals to convey information; however, it necessitates substantial exertion to acquire either complete knowledge or skill. Sign language recognition (SLR) has the intention to close the gap between the users and the non-users of sign language by identifying signs from video speeches. This is a fundamental but arduous task as sign language is carried out with complex and often fast hand gestures and motions, facial expressions and impressionable body postures. Nevertheless, non-manual features are currently being examined since numerous signs have identical manual components but vary in non-manual components. To this end, we suggest a novel manual and non-manual SLR system (MNM-SLR) using a convolutional neural network (CNN) to get the benefits of multi-cue information towards a significant recognition rate. Specifically, we suggest a model for a deep convolutional, long short-term memory network that simultaneously exploits the non-manual features, which is summarized by utilizing the head pose, as well as a model of the embedded dynamics of manual features. Contrary to other frequent works that focused on depth cameras, multiple camera visuals and electrical gloves, we employed the use of RGB, which allows individuals to communicate with a deaf person through their personal devices. As a result, our framework achieves a high recognition rate with an accuracy of 90.12% on the SIGNUM dataset and 94.87% on RWTH-PHOENIX-Weather 2014 dataset.

Keywords: CNN; CTC; recurrent neural network; sign language recognition; head pose
Mathematics Subject Classification: 37M10

1. Introduction

Sign language [1] is a visual and silent language accomplished with the kinetic movement of hand motions, facial expressions and body posture. Sign language represents an efficient and useful method of communication for both deaf individuals and individuals who have problems speaking in a regular tone of voice. Employing and understanding sign language demands a respectable amount of time,

apprenticeship and practice, which is not convenient and achievable for everyone. Furthermore, sign language has a large basis in culture [1,2], which also restricts its simplification potential. Even though computer vision and machine learning have reached a wide advancement in the past decade, it is still difficult to utilize sign language recognition (SLR), which automatically elucidates sign language and assists deaf-mute individuals in communicating with hearing individuals in their quotidian lives.

Compared with traditional action recognition, sign language recognition is a further exigent task. First, sign language demands both sensitive hand motions and total body gestures and clearly and precisely express its meaning. Moreover, facial expressions can be used to illustrate emotions. Analogous signs can even establish different meanings, which are reliant on the number of recurrences. Second, various signers may perform signs in a different way, which makes the recognition of sign language more challenging. Gathering many datasets from as many signers as possible is convenient yet pricey. Classical SLR systems principally prepare the dataset and manually use features, such as SIFT [3] and HOG [4], which are correlated with conventional classifiers such as SVM and KNN [5]. While deep learning is making major advancements, general methods for learning video and chronological series depictions (e.g., LSTM, RNN) and efficient video-based action recognition systems (e.g., 3D convolutional neural networks (CNNs)) are initially exploited for SLR assignments in [6,7]. Attention modules are joined with other modules to improve the precision to more adequately track down the information of local motion [8]. Additionally, [9] employs semantic segmentation and detection models to clearly lead the recognition network in a two-phase pipeline. Lately, body-based approaches have become suitable in gesture recognition tasks [10, 11] and define the growing attentiveness of their solid flexibility to the dynamic conditions and intricate background. Since the body-based approaches supply additional information to the RGB procedure, their whole results further enhance global achievement. Nonetheless, some insufficiency prevents their employment with the SLR method. Those body-based, deed recognition approaches depend on annotations of ground truth skeletons afforded by systems of gesture acquisition, thereby limiting themselves to publicly accessible datasets filmed in supervised surroundings. In addition, a large majority of motion acquisition systems only regard the coordinates of the principal body and does not supply a real observations of hands. As mentioned earlier, the data consists of inadequate information to handle SLR since signs are based on dynamic hand gestures and motions interrelated with different body parts. In [12], the authors tried to obtain information regarding various hand poses and skeletal structures by employing segregate models; their work suggested the use of an RNN-based model for SLR. However, their acquired hand poses were doubtful and the pattern could not correctly model the dynamics of the skeletons.

Head pose estimation is an influential way to convey additional information. Considering this, the main contributions of this paper are as follows:

1) Two features are disclosed, which are the anisotropic feature and the unsmooth variation feature. Inspired by the work of [13], a learning model of anisotropic angle distribution for the estimation of head poses is suggested. By employing a covariance pooling layer to apprehend the frame features of the second order, model learning is performed through an end-to-end CNN.

2) The suggested end-to-end adjoining model that combines both manual and non-manual features for SLR revealed substantial refinement in the accuracy performance for two publicly available datasets.

3) The suggested multimodal temporal representation (MTR) unit uses temporal receptive fields of various scales and presents a considerable enhancement in the concluding recognition achievement.

2. Related work

Over the past few years, SLR has reached important advancements and has acquired a high recognition rate through the improvement of convenient deep learning structures and the thrusting of computational potency [14–18]. There are some residual defiances in SLR, which are summarized in the simultaneous capture of overall body motion information, facial expressions and hand gestures. Authors in [19] suggested a multi-modal and multi-scale system that utilized spatial features at specific spatial ranges. An auto-encoder framework with a connectionist-based recognition component was suggested in [20] to model the sequence. Authors in [21] presented an end-to-end incorporation of a convolutional module within a hidden Markov model, and illustrated the approximation results in a Bayesian network. Authors in [8] suggested a CNN correlated with the attention component, which masters the spatio-temporal attributes from an unrefined video. In [22], the authors consolidated temporal convolutions and bidirectional recurrences with each other, which showed the efficiency of temporal information in gesture-based methods. In [23], the authors modeled a hierarchical attention network (HAN) with latent space to eliminate the temporal segmentation preprocessing. Nonetheless, these methods principally envisage raw visual features, which could be more effective to explicitly exploit various hand gestures and body movements. Authors in [24] presented a pose-based, temporal graph convolution network (GCN) that designs spatio-temporal reliances in trajectories of human posture. Authors in [25] adopted a hierarchical-LSTM auto-encoder pattern with visual content and a gloss incorporation for translation. They tackled various granularities by transmitting spatio-temporal transitions between frames. However, these methods were inefficient enough to exploit the total information of motion. Non-manual-based gesture recognition principally concentrates on examining peculiar patterns of motion and human joint position. Non-manual data can be used separately to carry out effective gesture recognition [26, 27]. Furthermore, it can as well be associated with other cues to obtain multi-cues learning desired for elevated recognition rates [28]. Recurrent neural networks are common for designing non-manual data, as is seen in [26, 27]. Newly, [29] is the first study to design a graph-based framework, named ST-GCN, for modeling the dynamic patterns in non-manual data through a GCN. This method attracts plenty of interest and a few ameliorations have been developed, such as in [30]. Especially, authors in [31] suggested an AS-GCN to delve into the latent joint connections to reinforce the achievement of recognition. Authors in [32] suggested a ResGCN, which adapts a bottleneck hierarchy from ResNet [33] to decrease parameters while growing the model's capability. Nonetheless, non-manual-based SLR systems have not been explored enough. In [34], the authors tried to directly spread out STGCN to SLR; however, the results were unsuccessful, and only reached about 60% recognition on 20 classes of sign language, which is unfortunately less than traditional approaches. The multi-cues method aims to examine gesture data received from either various devices or resources to boost the final achievement. This method is based on the hypothesis that various cues contain single motion information which could possibly complement each other and ultimately acquire particular and comprehensive action illustrations. For obtaining robust illustrations for downstream jobs, a view-invariant illustration learning framework was suggested in [35]. Authors in [36] deployed a shared weights network on a multi-cue script for obtaining cue vision for image classification. In [37], the authors proposed DA-Net, which is a view-independent and view-specific module for acquiring features and successfully combined the prediction scores together. In [22], the authors suggested a feature factorization framework that investigated the

specific information view shared for RGB-D gesture recognition. A cascaded residual auto-encoder was modeled in [38] to handle insufficient view classification settings. Inspired by the achievement of those multi-cue approaches, we intend to delve into more visual, gestural and hand cues alongside acquiring features from all appearances and combine them through a common framework to reach a more significant achievement.

3. Proposed approach

The global structure of the end-to-end, continuous SLR system suggested in this work shown in Figure 1. The pattern entails two-stream convolutional networks. The first ntwork aims to detect the head pose, while the second network includes the following three components: a spatial feature extraction component, a temporal feature extraction component and a multi-stage connectionist temporal classification (CTC) loss training component. In our model, there are five stages of gloss features. As the first step, the maximum a posteriori (MAP) estimation is used to design the network, which estimates the head pose. This network entails a convolutional pooling layer, a covariance pooling layer, and an output layer. For the second step, we employ the Resnet and two fully connected layers to the input sign language video to obtain the first-stage gloss features (spatial features). Hereafter, the temporal features are extracted by the suggested MTR unit. Specificaly, the spatial features proceed the prime MTR unit to obtain the second-stage gloss features; these latter features are successively adopted as the entry of the second MTR unit to obtain the third-stage gloss features, which become specified as the fourth-stage gloss features to the transformers timing coding. Lastly, the obtained five gloss features are combined and trained for model optimization by employing multi-stage CTC loss, and the conclusive SLR results are acquired by employing the fifth-stage gloss features.

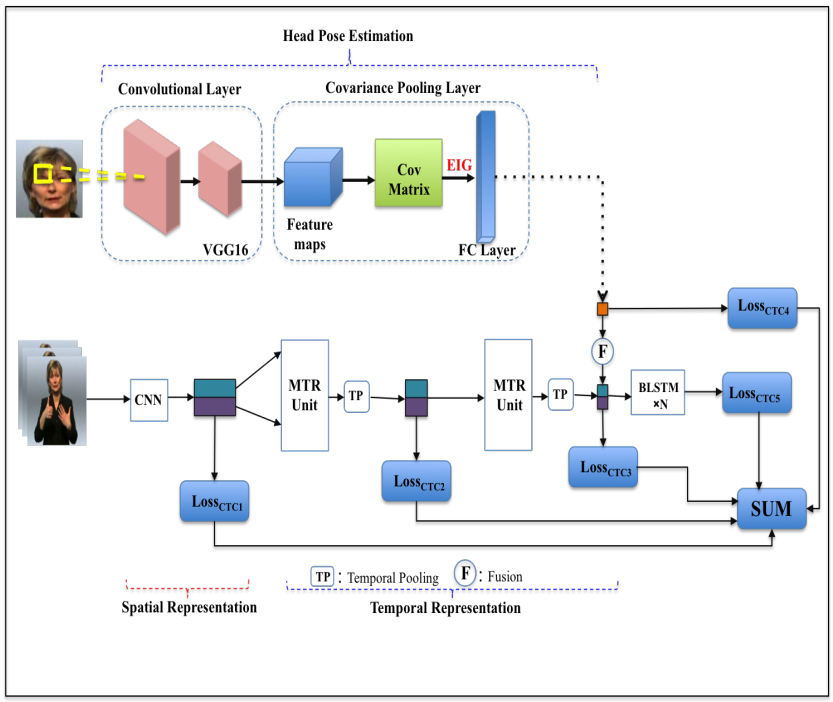

Figure 1. An overview of the proposed framework.

3.1. Head pose estimation

Head pose estimation signifies that the computer resolves the parameters of attitude and the position of the head in 3D space by examining and divining either the video sequence or the input images. Usually, the head pose is examined as a transformation of the inflexible body part. Head pose estimation works by measuring the 2D Euler angles, which incorporates the angles of yaw and pitch. Given a head pose angle Y and an input face image X, the occupation of the head pose estimation network uncovers the correct label Y from image X.

Two vectors instanced from the last fully connected layers are employed to calculate the similarity of the cosine. Given two frames, X1 and X2, the neural network is considering as a function that generates a vector of features. The formula is specified as follows:

$$FS(X_1, X_2) = Similarity(F(X_1), F(X_2)) = \frac{NN(X_1).NN(X_2)}{||NN(X_1)|| \times ||NN(X_2)||}. \tag{3.1}$$

The feature resemblances are computed between X_1, which represents the central position, and its adjoining positions are X_2, X_3, X_4 and X_5, respectively. All matrixes of resemblance are plotted and can be adjusted with 2D Gaussian distribution (Figures 2(b) and 4(c)) [13]. Next, the map scale can be obtained by computing all matrices of resemblance.

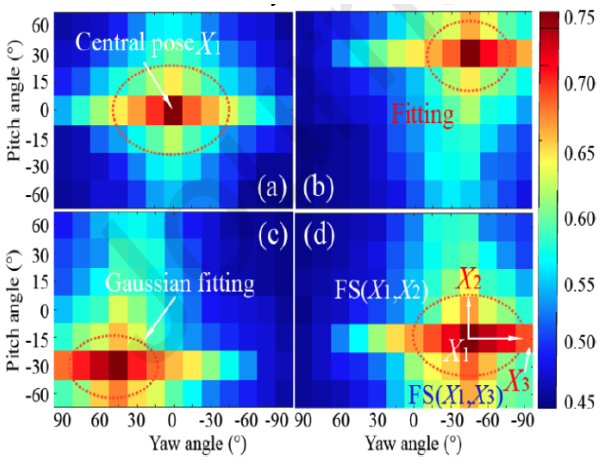

Figure 2. Anisotropic label construction.

Figure 3 shows that all poses of the head are arranged in a matrix. This latter has nine columns and 13 rows. Given a frame X of the head pose, its angle of axial pose is interpreted as $y_{\hat{m}\hat{n}} = (\hat{m}, \hat{n})$, where $\hat{n}$ and $\hat{m}$ are the column and row numbers of the pose image, respectively. The angle distribution $\hat{y}$ is interpreted as,

$$\hat{y} = \frac{g(y_{\hat{m}\hat{n}})}{\sum_m \sum_n g(y_{\hat{m}\hat{n}}} \tag{3.2}$$

and

$$g(y_{\hat{m}\hat{n}}) = \frac{1}{2\pi\sqrt{|\Omega|}} exp(-\frac{1}{2}((m-\hat{m})^2 + (n-\hat{n})^2)\Omega^{-1}), \tag{3.3}$$

where n represents the column number and m represents the row number in the matrix.

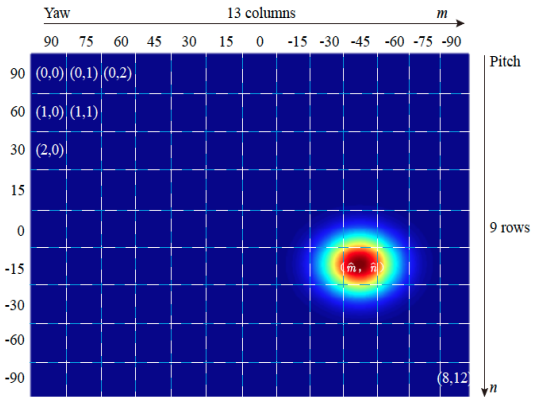

Figure 3. Head pose generation from Gaussian distribution.

Equation (3.3) shows that the appearance of the Gaussian distribution will be isotropic. The distribution appearance will be anisotropic whether the diagonal constituents are unequal or not. η is defined to obtain the 2D anisotropic Gaussian distribution, which can depict the anisotropic characteristic for the head pose estimation occupation. Based on the quantitative computation shown in Figure 3, the values of η are included in the range of (0.6, 1). In Figure 4, the property of unsmooth variation (i.e., when the angle range raises up, the image variations boost at first and then decline in the angle direction of the yaw) is transformed into the various standard deviation values σ of matrix M. Figure 4(a) depicts the angle distribution when the yaw = 0° and the pitch = 0°. Figure 4(b) depicts the angle distribution when the yaw = -45° and the pitch = 0°. We can note that the value of $\sigma 3$ is less than $\sigma 1$ and greater than $\sigma 2$.

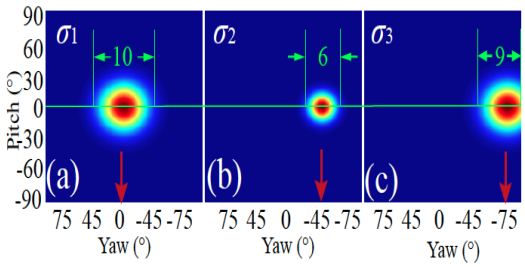

Figure 4. Unsmooth variation of head pose angle distribution.

3.2. Spatial feature representation

The spatial feature extraction module incorporates a leading network feature catcher and two fully connected layers.

As an entry video sequence, $VS = (vs_1, vs_2, ..., vs_T) = vs_t||_1^T \in \mathbb{R}^{T\times c\times h\times w}$ consists of T frames, where vs_t represents the t^{th} frame in the sequence, c depicts the channels number (c = 3) and $h * w$ represents the dimension of vs_t. VS is fed into the Resnet network, R_n, to acquire the feature composition $fc_1 = R_n(VS) \in \mathbb{R}^{T\times c_1}$; next, two fully connected layers are used to acquire the feature composition $fc_2 = R_{fc}(fc_1) \in \mathbb{R}^{T\times c_2}$, which represents the concluding spatial feature vector and the spatial feature with settled sizes. In this work, we have specified the concluded vector as the first-stage gloss feature. The dimensions of c_1 and c_2 are 512 and 1,024, respectively. The main role of using two fully connected layers next to the principal network is to incorporate features in the maps of the frame features that

have proceeded via numerous convolutional and pooling layers to obtain the high-stage significance of the frame features. We have applied a stochastic gradient stopping [39] between the Resnet and the fully connected layer to decrease the RAM usage and to speed up the training of the model.

3.3. Temporal feature representation

The temporal feature extraction module suggested in this work contain the following: An MTR unit and transformers. After passing through the MTR unit, gloss features of the second and the third stages are acquired. Ultimately, the gloss features of the fourth-stage are acquired after employing the transfomers.

1) MTR unit: These last years, exceptional continuous sign language recognition systems have been developed, though most of them utilize local features from the receptive fields of the designated temporalities. In sign language acquisition, the lengths of video sequences which represent various glosses are different. Additionally, the expertise of SL by various signers and certain other interferences over the filming operation produced incoherence in the length of the same word. Therefore, the obtained results will not be precise, thereby disturbing the achievement of temporal modeling. Figure 5 shows the proposed MTR unit in this work, which employs various ratios of temporal receptive fields to enhance the temporal representation efficiency. The MTR unit principally contains a multi-scale feature extraction and feature merging. Numerous one-dimensional CNNs with diverse convolution kernels are collaterally linked to make a multi-scale feature extraction element. The network is depicted as follows:

$$Net(t) = w(t) \times fc_2(t) = \sum_{S}^{i=0} fc_2 w(i)(t-i), \tag{3.4}$$

where $fc_2 \in \mathbb{R}^{T\times c_2}$ represents the weight, and $fc_2 \in \mathbb{R}^{T\times c_2}$ represent the obtained data, S represents the kernel size, $fc_2 \in \mathbb{R}^{T\times c_2}$ and T represents the length in terms of time. For the first-stage gloss feature, the feature size is initial updated from $fc_2 \in \mathbb{R}^{T\times c_2}$ to $fc_2^T \in \mathbb{R}^{c_2\times T}$. Afterward, it passes via the multi-scale feature extraction module. The multi-scale 1D-CNN has the equivalent number of channel dimensions number and various kernel sizes. The features number and the timing size make no changes over the treatment of feature pulling out. The kernel size of the beginning convolution layer is 2×2, while taking into account that the maximum size is S and the stride is two:

$$fc_2' = cat(Net_n(t)), \tag{3.5}$$

where fc_2' represents the exit that follows the multi-scale network and n represents the number of 1D-CNNs. Afterwards, we employed a feature merging and sub-sampling twice using a 2D-CNN, and $fc_3 \in \mathbb{R}^{c_2\times T_1}$ to obtain the second-stage gloss feature, where $T_1 = \frac{T}{2}$. We repeat this operation to procure the third-stage gloss feature $fc_4 \in \mathbb{R}^{c_2\times T_2}$, $T_2 = \frac{T_1}{2}$.

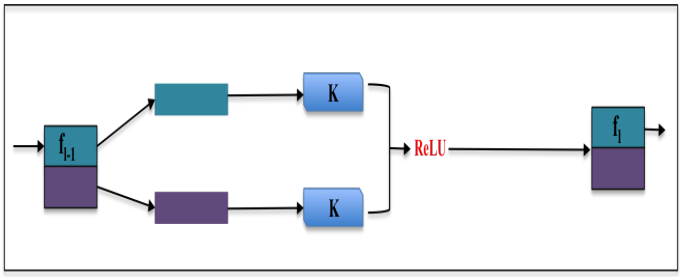

Figure 5. The input and output of the entity inside the multimodal temporal representation unit.

2) Transformers encoding: The transformers pattern is a traditional pattern of natural language processing introduced by Google in 2017. Instead of using the sequential structure of RNNs, it employs the self-attention mechanism, thereby permitting parallel training of the model and acquiring global information. The temporal sequence is encoded by employing the transformers encoding component after obtaining the temporal feature vector by the MTR unit, which leads to more precise temporal features. In our suggested model, two equivalent transformers encoding components were employed for the second CNN stream. The transformers encoding component was composed of a fully connected feed-forward element and a multi-head self-attention element. As the input of the multi-head self-attention element, we have introduced the third-stage gloss feature $fc_4 \in \mathbb{R}^{c_2 \times T_2}$ in parallel with the analogous position information. Afterward, the same process was iterated to acquire the final temporal feature $fc_5 \in \mathbb{R}^{c_2 \times T_2}$ via the temporal feature $fc'_4 \in \mathbb{R}^{c_2 \times T_2}$, which was acquired by the fully connected feedforward element; this gives us the fourth-stage gloss feature. In addition to the model's capacity to concentrate on various positions, multi-head self-attention also improves the capability of the attention structure to manifest the aspects among words inside the concerned sequences. In comparison with the self-attention of a single-head, every head in multi-head self attention preserves its own matrix (i.e., M_1, M_2, M_3) to accomplish distinct linear transformations in order, where every head further has its own particular meaningful information. Furthermore, the fully connected feedforward element consolidates the illustration in a non-linear manner, thereby permitting the features to be more eloquent.

3) Multi-stage CTC loss: Continuous SLR resides in faintly supervised learning. The entry is an unsegmented video sequence and misses a stringent accordance between labeled sequences and video frames. By succeeding to the step of the encoding of the entry sequence, it is highly suitable to employ CTC functioning as a decoder. The latter was initially conceived for recognizing speech, principally to carry out end-to-end temporal classification of the unsegmented signal to figure out the issue of contrasting lengths of entry and exit video sequences. During the last few years, CTC has been frequently employed in CSLR. It proposes a blank label $\{-\}$ to indicate labels that do not classified over-decoding (i.e., each word in the entry video sequence that does not apply to the vocabulary of sign language). Thereby, the entry and exit video sequences can be paired, and the dynamic programming algorithm can be employed for decoding [40]. Given an entry video sequence VS of T frames, every frame label is depicted by $\pi = (\pi_1, \pi_2, ..., \pi_T)$, where $\pi \in \nu \cap -$, and ν represents the vocabulary of sign language. The label posterior probability is given as follows:

$$p(\pi \| VS) = \Pi_{t=1}^{T} p(\pi_t \| VS) = \Pi_{t=1}^{T} Y_{t,\pi_t}, \tag{3.6}$$

for a specific sequence-stage label $s = (s_1, s_2, ..., s_L)$, where L represents the sequence word number.

CTC specifies a mapping that numerous instances of this entity are mapped to one instance of another entity, the process of which is to eliminate any duplicate and blank labels in the path. Therefore, the label conditional probability s is defined as the addition of the occurrence probabilities of all correlating paths:

$$p(s||VS) = \sum_{\pi \in B^{-1}(s)} p(\pi||VS), \tag{3.7}$$

where $B^{-1}(s) = \pi||B(\pi) = s$ represents the inverse mapping. A CTC loss is specified as the negative log-likelihood of the label conditional probability.

$$L_{CTC} = -\ln p(s||VS). \tag{3.8}$$

Therefore, the multi-stage CTC loss can be denoted as follows:

$$L_{sum} = -\ln \Pi_{i=1} n p(s||VS_i), \tag{3.9}$$

where n represents the CTC number.

The softmax function was implemented for normalization right after getting the four-stage gloss feature. The normalized outcome is decoded by CTC to acquire L_{CTC5}. Evenly, corresponding L_{CTC1}, L_{CTC2}, L_{CTC3}, and L_{CTC4} are acquired for the first, second, third, and fourth gloss features, respectively. Finally, these five CTC losses are summed to obtain the concluding loss for training:

$$L_{sum} = -\ln \Pi_{i=1} 4 p(s||VS_i). \tag{3.10}$$

4. Experiment and analysis

4.1. Experimental result and analysis

In this work, the suggested pattern of CNN (MNM-SLR) and another derivative of CNN (VGG16) [27] was inspected for SLR on two large-scale sign language benchmarks: SIGNUM and RWTH-PHOENIX-Weather 2014. In this division, the experimental results for these two patterns are debated, while, considering that a similar analysis with different state-of-art methods will be introduced in Section 4.2. Several metrics, such as processing time, loss, accuracy, and recognition prediction results, are employed to evaluate the performance of these two models.

4.1.1. Accuracy

To measure the classifier efficiency, the classification accuracy is the best used metric indicator. It is specified as the proportion of properly guessed instances to the overall number of instances in the dataset, as the following equation shows:

$$Accuracy = \frac{TN + TP}{FN + FP + TN + TP'}, \tag{4.1}$$

where TP, FP, TN, and FN are the true positive, false positive, true negative, and false negative, respectively. The accuracy of classification for SIGNUM employing MNM-SLR and VGG-16 is presented in Table 1. The concluding precision reached by the MNM-SLR model for continuous signs

is 88.96% and 94.37% for SIGNUM and RWTH-PHOENIX, respectively. The reached precision of classification using the VGG-16 model is 88.17% and 92.45% for SIGNUM and RWTH-PHOENIX, respectively. The results disclose that a higher achievement is obtained with MNM-SLR in comparison to VGG-16. Moreover, the performance of the two models has been tested on the expanded dataset. This was performed for the generalization of trained models. Data expandation is the operation of producing further data by transforming the initial possessed dataset. In this work, two supplementary instances per sample were produced by adopting the process of scaling and rotation. Therefore, a random inbound and outbound scaling of [0.7–1.4] and a random rotation in the interval [-15°, + 15°] were employed. The results of the classification for the expanded dataset are presented in Table 2. The distinguished augmented dataset results are sufficiently persuasive to demonstrate the generalization capability of the trained models.

Table 1. Accuracy and loss performance.

Model		MNM-SLR	VGG-16
Acc. (%)	SIGNUM	88.96	88.17
	RWTH-PHOENIX	94.37	92.45
Loss	SIGNUM	0.72	0.87
	RWTH-PHOENIX	0.53	0.64

Table 2. Results of classification for the expanded dataset.

Model	SIGNUM		RWTH-PHOENIX	
	Original	Aug	Original	Aug
MNM-SLR	88.96	90.12	94.37	94.87
VGG-16	88.17	89.11	92.45	93.42

4.1.2. Loss

In this work, the cross-entropy loss function is adopted to compute the loss that takes place in the multiple gestures classification of sign language, which is defined as follows:

$$Loss = \sum_{i=1}^{n} O_i . \log \widehat{O}_i, \tag{4.2}$$

where $\widehat{O}_i$ represents the i^{th} value in the output of the model, O_i depicts the analogous purpose value, and n represents the scalar value number in the exit of the model. The loss value calculated for two various patterns is shown in Table 1. This computed loss for all various CNN patterns regularly decreases with the augmentation of the iteration for a while, then subsequently obtains a determined value. For the SIGNUM dataset, the average loss for MNM-SLR and VGG-16 decrease to 0.72 and 0.87, respectively. For the RWTH-PHOENIX-Weather 2014 dataset, the loss for MNM-SLR and VGG-16 decreases to 0.53 and 0.64, respectively.

4.1.3. Confusion matrix

In the interest of better assessing the suggested framework, an alternative performance metric termed the confusion matrix, is also determined in this work. This matrix recapitulates the properly and wrongly predicted words of every class; therefore, the recognition precision of every class can be excerpted from it. Figures 6 and 7 demonstrate the confusion matrices of the obtained results employing our MNM-SLR system, which is applied on 26 classes of the RWTH-PHOENIX-Weather 2014 dataset. A qualitative analysis of the manual and non-manual confusion matrices (Figures 6 and 7) demonstrate that by employing non-manual features, it is possible to accurately determine more classes, which were classified incorrectly when employing solely manual features. We note that non-manual features can be employed to support differentiate various signs from each other.

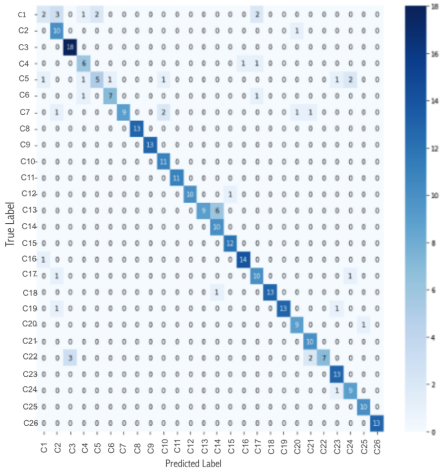

Figure 6. Confusion matrix with manual features only.

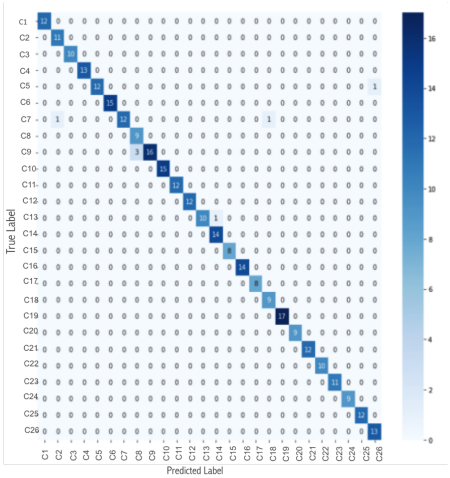

Figure 7. Confusion matrix with both manual and non-manual features.

4.1.4. Other parameters

Computational time is a crucial criterion for sign language recognition in real-time applications. The entire parameters for every convolutional layer can be represented as follows:

$$P_{conv} = (w \times h \times pf + 1) \times f, \tag{4.3}$$

where w and h represent the width and height of the filter, respectively, pf depicts the filter number of the previous layer and f depicts the filter number. The entire parameters for every fully connected layer (P_{fc}) can be represented as follows:

$$P_{fc} = (cl \times pl + 1) \times c, \tag{4.4}$$

where cl represents the current layer and pl represents the previous layer.

It is obvious by the distinction of the attainments that the suggested model of MNM-SLR employs a decreased computational time and fewer parameters as compared to other CNN architectures.

4.1.5. Cross-validation

K-fold is a cross-validation method to maintain the pattern achievement. Therefore, to assess the achievement on the whole data interval, a 10-fold cross-validation was employed for MNM-SLR. The assessment results for 10 folds are shown in Tables 3 and 4 for MNM-SLR and VGG-16, respectively.

Table 3. MNM-SLR framework results with 10-fold cross validation.

K-fold	SIGNUM	RWTH-PHOENIX-Weather
1	88.96	94.37
2	88.17	93.97
3	87.64	93.18
4	88.90	94.12
5	88.25	94.26
6	86.97	93.77
7	87.17	93.89
8	88.69	93.14
9	88.77	94.10
10	88.97	94.18

Table 4. VGG-16 framework results with 10-fold cross validation.

K-fold	SIGNUM	RWTH-PHOENIX-Weather
1	88.17	92.45
2	88.34	90.83
3	87.12	91.36
4	86.84	91.86
5	87.91	92.17
6	88.06	90.79
7	87.47	91.92
8	87.23	92.13
9	88.01	92.62
10	88.23	92.38

4.2. Comparaison with state of the art

DenseTCN is a dense temporal convolution network introduced by [41] and assumes the actions in hierarchical views. To learn the short-term correlation in this system, a temporal convolution (TC) is choosen in between neighboring features and extended to a dense hierarchical configuration. In [42], the authors nominated a CTM framework that enclosed the support of a temporal convolution pyramid module and a connectionist decoding pattern to design short-term and long-term sequence learning. Authors in [43] suggested a cross-modal learning model that weighed the text information for ameliorating vision-based CSLR. Hence, two efficient encoding networks are at first exercised for producing text and video enclosures before their alignment and mapping within a joint latent representation. Authors in [44] suggested a framework, namely ST-GCNs, which is an innovative deep-learning method that associates with spatio-temporal GCNs, which run on diverse, appropriately fusioned feature streams, assimilating signer's pose, motion information, appearance, and shape. The work of authors in [45] is sub-divided into three constituents: the first module is the feature extractor in a multi-view spatio-temporal Network (MSTN) that accurately extracts the spatio-temporal features of th RGB data and skeleton; the second module exemplifies an encoder network of SL based on the transformer, which can resolve dependency of long-term; the last module exemplifies a CTC decoder network. Table 5 exhibits that our proposed method obtains encouraging achievement, which is summarized by a decrease of the WER value to 30.7% on the RWTH-PHOENIX-Weather 2014 dataset. These results prove that the dynamic spatial correlation of SL sequences and the long-term temporal correlation can ameliorate learning of its visual features.

Table 5. Analysis of performance refinement on RWTH-PHOENIX-Weather 2014.

Methods	DEV		Test	
	del/ins	WER	del/ins	WER
ST-GCNs [44]	-	-	-	21.34
Dense TCN [41]	10.7 / 5.1	35.9	10.5 / 5.5	36.5
MSTN [45]	-	-	-	22.8
L_{joint} [43]	-	23.9	-	24.0
CTM [42]	11.6/6.3	38.9	10.9/6.4	38.7
STTN [46]	-	25.11	-	24.74
Our method	10.6/5.2	29.3	10.1/5.7	30.7

4.3. Ablation study

To evaluate the contributions of the designed model, we have performed an ablation study on the RWTH-PHOENIX-Weather 2014 dataset. An ablation study can analyze the different components that influence the performance of the system. As shown in Table 6, it is feasible to evaluate the impact of each proposed training structure. For this determination, the proposed model was trained either (i) with non-manual features or (ii) without non-manual features. As concluded from Table 6, the proposed model reveals a higher achievement when employing all modalities together, thereby yielding a 30.7% WER.

Table 6. Ablation study on RWTH-PHOENIX-Weather 2014. W/O NMF means without non-manual features and W NMF means with non-manual features.

Methods	DEV del/ins	DEV WER	Test del/ins	Test WER
W/O NMF	11.2/6.3	33.43	10.8/5.9	37.67
W NMF	10.6/5.2	29.3	10.1/5.7	30.7

5. Conclusions

In this work, we suggest an inventive training approach to produce a nominated feature extraction module, which was thoroughly employed to better understand the convenient sign language gloss on video sequences, while continuing to benefit from the iteratively cleansed alignment propositions. We advance a multi-modal method to integrate the head position and motion gestures from video sequences of sign language, which supplies superior spatio-temporal representations for gestures. The substantial contribution of the proposed work is its capacity to recognize complex signs. It demonstrates that by employing non-manual features, it is possible to accurately determine more classes, which were classified incorrectly when employed solely as manual features. It was affirmed via experiments that our MNM-SLR framework achieves a state-of-the-art performance on continuous sign langue recognition with an accuracy of 90.12% on the SIGNUM dataset and 94.87% on the RWTH-PHOENIX-Weather 2014 dataset.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research has been funded by Deputy for Research & Innovation, Ministry of Education through Initiative of Institutional Funding at University of Ha'il-Saudi Arabia through project number IFP-22 018.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. W. C. Stokoe, Sign language structure, *Annu. Rev. Anthropol.*, **9** (1980), 365–390. http://dx.doi.org/10.1146/annurev.an.09.100180.002053
2. J. Napier, L. Leeson, *Sign language in action*, London: Palgrave Macmillan, 2016. http://dx.doi.org/10.1057/9781137309778

3. D. Lowe, Object recognition from local scale-invariant features, *Proc. IEEE Int. Conf. Comput. Vision*, **2** (1999), 1150–1157. http://dx.doi.org/10.1109/ICCV.1999.790410

4. Q. Zhu, M. C. Yeh, K. T. Cheng, S. Avidan, Fast human detection using a cascade of histograms of oriented gradients, *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recogn.*, 2006, 1491–1498. http://dx.doi.org/10.1109/CVPR.2006.119

5. A. Memiş, S. Albayrak, A Kinect based sign language recognition system using spatio-temporal features, *Proc. SPIE Int. Soc. Opt. Eng.*, **9067** (2013), 179–183. http://dx.doi.org/10.1117/12.2051018

6. O. Sincan, H. Keles, Using motion history images with 3D convolutional networks in isolated sign language recognition, *IEEE Access*, **10** (2022), 18608–18618. http://dx.doi.org/10.1109/ACCESS.2022.3151362

7. G. Castro, R. R. Guerra, F. G. Guimarães, Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps, *Expert Syst. Appl.*, **215** (2023), 119394. http://dx.doi.org/10.1016/j.eswa.2022.119394

8. J. Huang, W. G. Zhou, H. G. Li, W. P. Li, Attention-based 3D-CNNs for large-vocabulary sign language recognition, *IEEE T. Circ. Syst. Vid.*, **9** (2018), 2822–2832. http://dx.doi.org/10.1109/TCSVT.2018.2870740

9. K. Lim, A. Tan, C. P. Lee, S. Tan, Isolated sign language recognition using convolutional neural network hand modelling and hand energy image, *Multimed. Tools Appl.*, **78** (2019), 19917–19944. http://dx.doi.org/10.1007/s11042-019-7263-7

10. M. Terreran, M. Lazzaretto, S. Ghidoni, Skeleton-based action and gesture recognition for human-robot collaboration, *Intell. Auton. Syst.*, **577** (2022), 29–45. http://dx.doi.org/10.1007/978-3-031-22216-0_3

11. L. Roda-Sanchez, C. Garrido-Hidalgo, A. S. García, T. Olivares, A. Fernández-Caballero, Comparison of RGB-D and IMU-based gesture recognition for human-robot interaction in remanufacturing, *Int. J. Adv. Manuf. Technol.*, **124** (2023), 3099–3111. http://dx.doi.org/10.1007/s00170-021-08125-9

12. W. Aditya, T. K. Shih, T. Thaipisutikul, A. S. Fitriajie, M. Gochoo, F. Utaminingrum, et al., Novel spatio-temporal continuous sign language recognition using an attentive multi-feature network, *Sensors*, **22** (2022), 6452. http://dx.doi.org/10.3390/s22176452

13. H. Liu, H. Nie, Z. Zhang, Y. F. Li, Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction, *Neurocomputing*, **433** (2020), 310–322. http://dx.doi.org/10.1016/j.neucom.2020.09.068

14. S. Sharma, R. Gupta, A. Kumar, Continuous sign language recognition using isolated signs data and deep transfer learning, *J. Amb. Intel. Hum. Comp.*, 2021, 1–12. http://dx.doi.org/10.1007/s12652-021-03418-z

15. O. Koller, S. Zargaran, H. Ney, R. Bowden, Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs, *Int. J. Comput. Vision*, **126** (2018), 1311–1325. http://dx.doi.org/10.1007/s11263-018-1121-3

16. O. Koller, H. Ney, R. Bowden, Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled, *IEEE Conf. Comput. Vision Pattern Recogn.*, 2016, 3793–3802. http://dx.doi.org/10.1109/CVPR.2016.412

17. O. Koller, S. Zargaran, H. Ney, R. Bowden, Deep sign: Hybrid CNN-HMM for continuous sign language recognition, *Brit. Conf. Mach. Vision*, 2016.

18. O. Koller, H. Ney, R. Bowden, Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs, *IEEE Conf. Comput. Vision Pattern Recogn.*, 2017, 4297–4305. http://dx.doi.org/10.1109/CVPR.2017.364

19. O. Özdemir, İ. Baytaş, L. Akarun, Multi-cue temporal modeling for skeleton-based sign language recognition, *Front. Neurosci.*, **17** (2023), 1148191. http://dx.doi.org/10.3389/fnins.2023.1148191

20. H. Butt, M. R. Raza, M. R. Ramzan, M. J. Ali, M. Haris, Attention-based CNN-RNN Arabic text recognition from natural scene images, *Forecasting*, **3** (2021), 520–540. http://dx.doi.org/10.3390/forecast3030033

21. P. P. Roy, P. Kumar, B. G. Kim, An efficient sign language recognition (SLR) system using camshift tracker and hidden markov model (HMM), *SN Comput. Sci.*, **2** (2021), 1–15. http://dx.doi.org/10.1007/s42979-021-00485-z

22. L. Pigou, A. Oord, S. Dieleman, M. V. Herreweghe, J. Dambre, Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video, *Int. J. Comput. Vision*, **126** (2018), 430–439. http://dx.doi.org/10.1007/s11263-016-0957-7

23. J. Huang, W. G. Zhou, Q. L. Zhang, H. Q. Li, W. P. Li, Video-based sign language recognition without temporal segmentation, *Proc. AAAI Conf. Artif. Intell.*, **32** (2018). http://dx.doi.org/10.1609/aaai.v32i1.11903

24. K. Han, X. Y. Li, Research method of discontinuous-gait image recognition based on human skeleton keypoint extraction, *Sensors*, **23** (2023), 7274. http://dx.doi.org/10.3390/s23167274

25. D. Wategaonkar, R. Pawar, P. Jadhav, T. Patole, R. Jadhav, S. Gupta, Sign gesture interpreter for better communication between a normal and deaf person, *J. Pharm. Negat. Result.*, 2022, 5990–6000. http://dx.doi.org/10.47750/pnr.2022.13.S07.731

26. M. Jebali, A. Dakhli, M. Jemni, Vision-based continuous sign language recognition using multimodal sensor fusion, *Evol. Syst.*, **12** (2021), 1031–1044. http://dx.doi.org/10.1007/s12530-020-09365-y

27. M. Jebali, A. Dakhli, W. Bakari, Deep learning-based sign language recognition system for cognitive development, *Cogn. Comput.*, 2023, 1–13. http://dx.doi.org/10.1007/s12559-023-10182-z

28. V. Choutas, P. Weinzaepfel, J. Revaud, C. Schmid, PoTion: Pose motion representation for action recognition, *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2018, 7024–7033. http://dx.doi.org/10.1109/CVPR.2018.00734

29. S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, *Proc. AAAI Conf. Artif. Intell.*, **32** (2018). http://dx.doi.org/10.1609/aaai.v32i1.12328

30. M. Bicego, M. Vázquez-Enríquez, J. L. Alba-Castro, Active class selection for dataset acquisition in sign language recognition, *Image Anal. Proc.*, 2023, 303–315. http://dx.doi.org/10.1007/978-3-031-43148-7_26

31. M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, *IEEE Conf. Comput. Vision Pattern Recogn.*, 2019, 3590–3598. http://dx.doi.org/10.1109/CVPR.2019.00371

32. Y. F. Song, Z. Zhang, C. Shan, L. Wang, Constructing stronger and faster baselines for skeleton-based action recognition, *IEEE T. Pattern Anal.*, **45** (2022), 1474–1488. http://dx.doi.org/10.1109/TPAMI.2022.3157033

33. Z. Wu, C. Shen, A. Hengel, Wider or deeper: Revisiting the ResNet model for visual recognition, *Pattern Recogn.*, **90** (2019), 119–133. http://dx.doi.org/10.1016/j.patcog.2019.01.006

34. N. Takayama, G. Benitez-Garcia, H. Takahashi, Masked batch normalization to improve tracking-based sign language recognition using graph convolutional networks, *IEEE Int. Conf. Autom. Face Gesture Recogn.*, 2021, 1–5. http://dx.doi.org/10.1109/FG52635.2021.9667007

35. Ç. Gökçe, Ç. Özdemir, A. A. Kındıroğlu, L. Akarun, Score-level multi cue fusion for sign language recognition, *Eur. Conf. Comput. Vision*, 2020, 294–309. http://dx.doi.org/10.48550/arXiv.2009.14139

36. L. Tarrés, G. I. Gállego, A. Duarte, J. Torres, X. Giró-i-Nieto, Sign language translation from instructional videos, *IEEE Conf. Comput. Vision Pattern Recogn. Work.*, 2023, 5625–5635. http://dx.doi.org/10.1109/CVPRW59228.2023.00596

37. O. Sincan, A. Tur, H. Keles, Isolated sign language recognition with multi-scale features using LSTM, *Proc. Commun. Appl. Conf.*, 2019, 1–4. http://dx.doi.org/10.1109/SIU.2019.8806467

38. Q. Guo, S. J. Zhang, L. W. Tan, K. Fang, Y. H. Du, Interactive attention and improved GCN for continuous sign language recognition, *Biomed. Signal Proces.*, **85** (2023), 104931. http://dx.doi.org/10.1016/j.bspc.2023.104931

39. Z. Niu, B. Mak, Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition, *Eur. Conf. Comput. Vision*, 2020, 172–186. http://dx.doi.org/10.1007/978-3-030-58517-4_11

40. A. Hao, Y. Min, X. Chen, Self-mutual distillation learning for continuous sign language recognition, *Int. Conf. Comput. Vision*, 2021, 11303–11312. http://dx.doi.org/10.1109/ICCV48922.2021.01111

41. D. Guo, S. Wang, Q. Tian, M. Wang, Dense temporal convolution network for sign language translation, *Int. Joint Conf. Artif. Intell.*, 2019, 744–750. http://dx.doi.org/10.24963/ijcai.2019/105

42. D. Guo, S. G. Tang, M. Wang, Connectionist temporal modeling of video and language: A joint model for translation and sign labeling, *Int. Joint Conf. Artif. Intell.*, 2019, 751–757. http://dx.doi.org/10.24963/ijcai.2019/106

43. I. Papastratis, K. Dimitropoulos, D. Konstantinidis, P. Daras, Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space, *IEEE Access*, **8** (2020), 91170–91180. http://dx.doi.org/10.1109/ACCESS.2020.2993650

44. M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, P. Maragos, Spatio-temporal graph convolutional networks for continuous sign language recognition, *IEEE Int. Conf. Acous. Speech Signal Proc.*, 2022, 8457–8461. http://dx.doi.org/10.1109/ICASSP43922.2022.9746971

45. R. Li, L. Meng, Multi-view spatial-temporal network for continuous sign language recognition, *Comput. Vision Pattern Recogn*, 2022. http://dx.doi.org/10.48550/arXiv.2204.08747

46. Z. C. Cui, W. B. Zhang, Z. X. Li, Z. Q. Wang, Spatial-temporal transformer for end-to-end sign language recognition, *Complex Intell. Syst.*, **9** (2023), 4645–4656. http://dx.doi.org/10.1007/s40747-023-00977-w

AIMS Press