GMR
The Original

A HYBRID HOG–PCA AND VISION TRANSFORMER–BASED SEQUENCE-TO-SEQUENCE FRAMEWORK FOR LOW-RESOURCE SENTENCE-LEVEL INDIAN SIGN LANGUAGE VIDEO-TO-TEXT TRANSLATION

Ashwini Nikhil Shinde[1],Dr. Rajeswari Kannan[2]

[1]Research Scholar, Pimpri Chichwad College of Engineering, Pune, India, shinde17ashwini@gmail.com, ORCID: https://orcid.org/0009-0006-4748-2555
[2]Professor, Pimpri Chinchwad College of Engineering, Pune, India, kannan.rajeswari@pccoepune.org, https://orcid.org/0000-0002-5392-3920

ABSTRACT
Automatic Indian Sign Language (ISL) video-to-text translation is still a problem because of the complex spatial-temporal relationships between gestures, signer variability and scarcity of annotated training material. A hybrid sequence-to-sequence model with HOG–PCA and Vision Transformer (ViT) for low-resource sentence-level ISL translation is proposed in this study. The proposed architecture consists of concatenating handcrafted gradient descriptors with the contextual embeddings obtained from the ViT. An autoregressive decoder produces English sentences from the temporal Transformer encoder, which models the sequential dependencies of the gestures. Comparative experiments, low-resource evaluations, architectural ablations and PCA dimensionality analysis were conducted to evaluate translation performance. The proposed framework achieved BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 30.80, 22.60, 17.40, and 13.20, respectively, alongside ROUGE-L and chrF scores of 37.60 and 32.80. Over the conventional ViT Transformer and the parameter-matched baseline, BLEU-4 gained 4.70 and 3.50 points, respectively. The hybrid model achieved better performance than the ViT-only model in all the training-data percentages with 128 PCA components showing the best translation performance. Based on these findings, the complementary handcrafted and deep visual representations can be considered as a potential approach for improving ISL sentence-level translation with limited supervision.

KEYWORDS: Indian Sign Language; Video-to-Text Translation; HOG–PCA; Vision Transformer; Sequence-to-Sequence Learning.

INTRODUCTION

Sign language is a key communication tool for Deaf and hard of hearing people, and communicates language using a combination of body position, spacing, facial expression and hand movements. But communication barriers can occur when sign language users communicate with people who do not know sign language. Automatic sign language translation offers a good solution to overcome these barriers by translating a continuous visual sign language into meaningful sentences. Recent developments in artificial intelligence, computer vision, and natural language processing have greatly facilitated the creation of translation systems with the ability to learn complex visual–linguistic relationships. However, sentence level translation is still difficult due to the aspects of space, time, and grammar being a part of sign languages and not being adequately conveyed by isolated gesture recognition. Shahin, Ismail (2024) pointed out the evolution of the recognition systems from the traditional ones to deep learning and Transformer-based, focusing on the role of contextual representation learning.

In contrast to sign recognition, when translating a sentence, the continuous sequence of gestures must be interpreted and coherent textual expressions must be produced. The task of temporal segmentation ambiguity, coarticulation, changes in signing speed, and dependency between consecutive signs. In the study by Tanzer et al. (2024), the authors showed that processing sentence-level videos independently can be problematic since they might not provide enough information to interpret individual signing segments. Therefore, it is necessary to consider the spatial information and the temporal relationships and at the same time, to maintain the linguistic meaning of the continuous signing sequences in the translation architectures.

The language of Indian Sign Language (ISL) poses further computational problems as a result of the comparatively low availability of annotated video–text resources. Joshi et al. (2023) presented ISLTranslate, which comprises about 31,000 sentence and phrase pairs in ISL and English languages, thus contributing to the research on continuous translation. However, there are still challenges for reliable translation, such as signer variability, gesture complexity, the differences in the environment and limited training supervision. This was then followed by the work of Joshi et al. (2025) which demonstrated that linguistically-informed pose-sequence construction can aid in the better learning of translation. All these advances indicate that obtaining informative representations and effectively using available annotated data are important.

Sign language translation, including attention-based video and language sequence modeling, has made tremendous progress with the development of transformer-based methods. Camgoz et al. (2020) presented a joint recognition and translation model, which has shown effectiveness for continuous signing using Transformers. The more recent ones have

investigated visual–language pretraining and gloss-free translation to minimize the use of intermediate sign annotations. However, the quality of the visual representation, the complexity of the model, the temporal modeling and conditions of evaluation also influence the performance of the translation. Sincan et al. (2025) pointed out that unbiased benchmarking and controlled architectural comparisons should be considered when assessing translation systems.

Popular options for Conv are features, skeletons or embedding deeply in visuals. These techniques will be helpful for recording key elements of the signing, but individual representation mechanisms may not record supplementary information. Hand configuration involves using Histogram of Oriented Gradients (HOG) for representing the localized edges, contours and orientation distributions and Principal Component Analysis (PCA) for reducing the redundancies to obtain compact structural features. In contrast, Vision Transformers (ViTs) model relationships between spatial image regions in a context using self-attention. The combination of HOG–PCA descriptors with the ViT embeddings thus could be a way to fuse localized structural features with global contextual information. Yet the advantages of this integration, in particular for sentence level ISL translation in low resource settings have not been sufficiently proven by controlled comparisons.

To fill this research gap, the present paper presents a hybrid HOG–PCA feature based and ViT based sequence-to-sequence framework for continuous ISL video-to-English text translation. We propose a handcrafted gradient descriptors, PCA-based dimensionality reduction, and Transformer-derived visual representations by concatenating these features. Sequential dependencies are then modelled by a temporal Transformer encoder and English sentences are generated by an autoregressive Transformer decoder. The framework is assessed in contrast with HOG–PCA Transformer, ViT Transformer and ViT Transformer with same number of parameters. Further experiments compare the performance of translation in the cases where only a limited amount of training data is available, architectural ablations, feature-fusion, and PCA dimensionality variations.

The study addresses the following three research objectives:

- To develop a hybrid visual representation and sequence-to-sequence translation framework integrating HOG descriptors, PCA dimensionality reduction, ViT embeddings, temporal Transformer encoding, and autoregressive decoding for continuous ISL video-to-English text translation.
- To evaluate translation performance under different training-data conditions by comparing the proposed architecture with HOG–PCA Transformer, ViT Transformer, and parameter-matched ViT Transformer baselines using BLEU, ROUGE-L, and chrF evaluation metrics.
- To investigate the contributions of individual architectural components through systematic ablation experiments and PCA dimensionality analysis, examining the effects of handcrafted features, contextual embeddings, temporal modeling, and feature-fusion strategies.

The main contribution of this study is the combination of complementary structural features (HOG–PCA) and contextual features (ViT) in a single temporal sequence-to-sequence translation framework. As opposed to methods which are based only on handcrafted or deep visual features, the proposed framework is explored as a combination of handcrafted and deep visual features under varying amounts of annotated training data. The reported experimental results show that the hybrid architecture can obtain a BLEU-4 score of 13.20 compared to 8.50 of the conventional ViT Transformer and 9.70 of the parameter-matched configurations. The proposed model also achieves ROUGE-L and chrF scores of 37.60 and 32.80, respectively. These findings confirm that in the evaluated conditions, complementary visual representations for the improvement of sentence-level ISL translation have potential.

This article follows with related literature, methodology, experimental set-up, results, discussion and conclusions.

2. LITERATURE REVIEW

2.1. Evolution of Automated Sign Language Translation

From the traditional sign language recognition systems, automated sign language translation has progressed to deep learning-based systems that can handle complete sequence of sign language and output a natural language sentence. Most early computational methods were geared towards the task of individual sign recognition based on carefully designed visual features and sequential recognition models. These techniques allowed the identification of individual signs, but lacked the capacity to represent complex relationships in language. Camgoz et al. (2020) proposed Sign Language Transformers, which are based on an attention model for recognizing and translating Sign Language simultaneously. Their work proved to be successful in showing how sequence modeling works to determine relationships between sequences of signs and textual representations. De Coster et al. (2020) investigated further the use of Transformer networks for sign language recognition, which supports the use of attention mechanisms for sequential visual data.

Even with these, sentence-level translation is more challenging than isolated sign classification as boundaries and context of gestures, as well as linguistic and semantic dependencies are not explicitly stated in continuous videos. Tanzer et al. (2024) pointed out some drawbacks of sentence-level task formulations if the interpretation of a sentence requires a wider context of the discourse. The results suggest that visual sequence representation and contextual language generation are both important issues that should be tackled in translation architectures.

2.2. Indian Sign Language Datasets and Low-Resource Translation

Annotation of datasets has played a pivotal role in the advancement of Indian Sign Language processing. Joshi et al. (2023) released ISLTranslate containing about 31,000 ISL–English sentence and phrase pairs; they set up the baseline translation experiments with Transformer-based models. This contribution set the bases for ongoing ISL translation research. However, challenges of signer variability, sentence complexity and the lack of supervision for various linguistic structures remain when the annotated data is available.

More recently, Joshi et al. (2025) suggested a linguistically-inspired pose-sequence pretraining strategy, named PoseStitch-SLT, which aims to enhance translation using examples generated via templates. They found in their assessment that there is a potential of further structured supervision for improving gloss-free translation in a situation where data is limited. The approach also revealed the significance of effective representation learning in the context where it is challenging to get large collections of sentence-aligned sign language video.
These studies show that there still are ongoing advancements in ISL translation and that other means of visual representation should be explored. Especially, combining small handcrafted descriptors and large contextual embeddings is worth exploring as an alternative to better learning from small amounts of annotated training data.

2.3. Deep Learning and Gloss-Free Translation Architectures

By enabling the visual representation learning and cross-modal sequence alignment, deep learning architectures have played a significant role in continuous sign language translation. However, to avoid such explicit gloss supervision for gloss-free translation, Yin et al. (2023) proposed gloss attention, which can capture semantic relationships and local contextual boundaries through attention mechanisms. Their method coped with the challenge of directly learning useful temporal structure from continuous signing videos.
Zhou et al. (2023) also proposed a visual–language pretrained gloss-free translation framework. Their architecture increased the accuracy of the alignment between signing sequences and target sentences by borrowing both visual and linguistic information. Hamidullah et al. (2024) investigated using target-language semantic representations for learning translation and the utility of sentence embedding supervision as an alternative to gloss annotation.
More recently, efforts have been made to integrate LLM in SL translation. Gong et al. (2024) proposed SignLLM for converting sign videos into language representations via visual tokenization and cross-modal alignment. Kim et al. (2025) created a multimodal large language model (LLM) based approach, which combines textual description of sign language components with visual–linguistic representation learning. While these methods show the promise of language models pre-trained with external tools for representation learning, they also add extra layers of architectural complexity and reliance on external resources.
In sum, all these studies emphasize the need for discriminative visual features, temporal relationships and visual-linguistic alignment. They do not, however, directly guarantee the measurable merits of the PCA compressed HOG descriptor and embedding from the ViT network for sentence-level ISL translation.

2.4. Handcrafted Visual Features and Vision Transformer Representations

The visual feature extraction is pivotal in sign language translation as there are subtle hand orientation, shape, position and spatial relationships in sign language. To visualize the local gradient distributions from a 2-D image, Dalal and Triggs (2005) proposed to use Histogram of Oriented Gradients, which to date is a popular method for visualizing local structure in a 2-D image. HOG features extracting the information about the edges and orientation of the hand is localised and can be additional information to describe the hand configurations in signing videos.
But, high dimensional hand crafted descriptors might include similar and redundant information. Principal Component Analysis is a statistical technique that can be used for dimensionality reduction, and maintain the main directions of variation of features (Jolliffe, 2002). It is then highly valuable to feed the HOG with the PCA and get a small representation of the local structural information, but one should consider the number of components to retain as it depends on the application and training data.
Complementary approach is deep visual representation learning. Recently, Vision Transformers were shown to be able to treat images as a sequence of patch embeddings, and to learn these relationships through self-attention by Dosovitskiy et al. (2021). In contrast to fixed gradient descriptors, ViT representations have complex spatial interactions between the different regions of the images that are learned from data. Still, there are differences in the visual information which are highlighted in the embedding space of ViT and the handcrafted descriptors. They can thus be used in combination to give complementary structure and context, but the success of this combination has to be shown in experiments.

2.5. Temporal Modeling and Sequence-to-Sequence Translation

In continuous sign language videos, the sign language is temporally dependent and the meaning of the individual movements cannot necessarily be derived from the individual frames. To this end, temporal encoding of successive gestures, necessary to capture relationships between them, and to hold information relevant to sentence generation is required. Transformer (Vaswani et al., 2017) is an architecture that uses self-attention to capture relationships between elements within the sequence, but not necessarily requires recurrent computation.
Transformer encoders can capture relationships between a series of temporally ordered visual embeddings, and autoregressive decoders can produce tokens in the target language based on the encoded information. This encoder–decoder framework offers a good computational foundation for continuous sign language video-to-text translation. Camgoz et al. (2020) showed its importance in sign language processing for Joint Recognition and Translation.
More recent studies have taken into account temporal visual representations efficiency. To overcome the longer visual sequence and the low efficiency of tokens, Low et al. (2025) presented a segment-aware gloss-free encoding (SAGE) method. They present their results, which demonstrate the computational consequences of having long signing sequences and the relevance of choosing informative visual representations.
In this direction, a complementary spatial description will be integrated beforehand the temporal Transformer encoding, and it will be analyzed whether the complement result in improving the sentence-level translation performance by using increasingly complex language model or not.

2.6. Research Gap and Positioning of the Proposed Framework

From the surveyed literature, we see that the fields of Transformer-based translation, learning without glosses, visual–language pretraining, pose-based supervision and efficient sequence modeling have progressed. However, three issues related to this have led to the current investigation. First, there is a lack of proper analysis of how the fusion of handcrafted gradient descriptor (based on PCA) and contextual representations (based on ViT) for continuous ISL translation works. Second, it is not clear how beneficial these hybrid representations can be in scenarios with systematically limited training data. Third, controlled comparisons have to be done to distinguish whether improvement is due to complementary feature information or due to increased capacity of the model or to the differences in the architecture.

To solve the problem, we propose a hybrid framework that integrates localized structural features (HOG) and contextual visual features (ViT) before temporal Transformer encoding, dubbed as HOG–PCA–ViT sequence-to-sequence framework. The handcrafted-feature, ViT-only and parameter-matched ViT architectures are compared. Low-resource experiments study the impact of having or not having training data and ablation and PCA dimensionality analyses study the impact of individual components.

The key research innovation is not the use of HOG or PCA or ViT or Transformer-based decoding alone, but their combination and systematic analyses for ISL sentence-level video-to-text translation. The study explores the complementary nature of handcrafted and deep visual representations and aims to offer empirical results on hybrid feature learning with limited annotated supervision.

3. Proposed Methodology

3.1. Research Design and Framework

The approach used in this study was quantitative experimentation, in which a hybrid Histogram of Oriented Gradients–Principal Component Analysis (HOG–PCA) and Vision Transformer (ViT) based sequence to sequence framework for sentence level Indian Sign Language (ISL) video to text translation was developed and evaluated. In the proposed architecture, handmade spatial descriptors were leveraged to complement deep contextual representations of the visual data to improve the translation performance in the limited annotated training case. It involves three key steps: pre-processing the video, extracting features using HOG–PCA, representation learning using ViT, hybrid feature fusion, temporal Transformer encoding, and autoregressive English sentence generation. Comparative evaluation, low resource experiments, architectural ablations and PCA dimensionality analysis were used to compare the translation performance and the contribution of individual components.

The sequence-to-sequence framework that we propose, based on HOG–PCA–ViT, is presented in Figure 1. The architecture is composed of two parallel feature extraction layers (HOG–PCA and ViT), feature concatenation, temporal Transformer encoding and autoregressive English sentence generation.

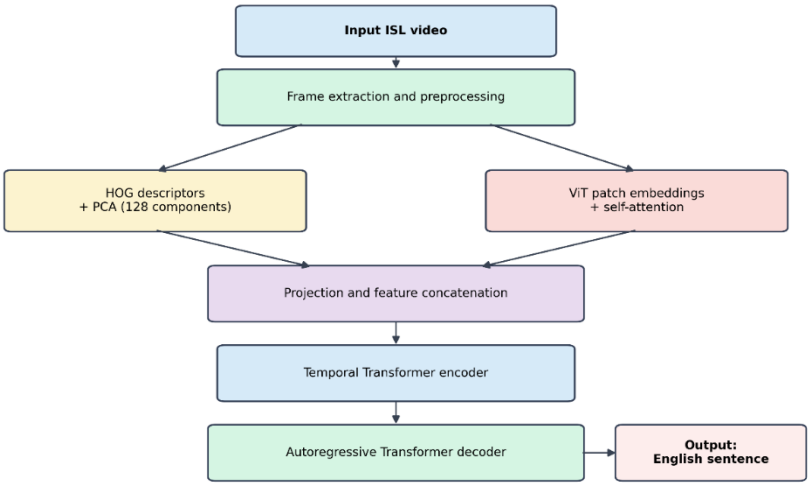


Figure 1. Proposed hybrid HOG–PCA and Vision Transformer-based sequence-to-sequence framework for sentence-level Indian Sign Language video-to-text translation.

3.2. Dataset and Video Preprocessing

The iSign benchmark, introduced by Joshi et al. (2024), provides annotated Indian Sign Language video–text pairs suitable for continuous sign language translation. Each sample consists of a signing video and its corresponding English sentence. The dataset was mathematically represented as:

$$D = \{(V_i, Y_i) \mid i = 1, 2, \ldots, N\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where D denotes the dataset, V_i represents the ith ISL video, Y_i indicates its corresponding reference English sentence, and N represents the total number of video–text pairs.

Video preprocessing involved frame extraction, temporal sampling, spatial resizing, and pixel normalization. Each sampled video sequence was expressed as:

$V_i = (I_{i,1}, I_{i,2}, \ldots, I_{i,t})$ ..... (2)

where $I_{i,t}$ denotes the frame at temporal position t, and T represents the total number of sampled frames. Temporal ordering was preserved to maintain gesture continuity. The evaluation protocol distinguished training, validation, and testing partitions, with training data supporting parameter estimation and validation data supporting model selection.

3.3. HOG–PCA Feature Extraction

The handcrafted feature extraction branch utilized HOG descriptors to encode localized edge, gradient orientation and structural properties of signing gestures. Based on Dalal and Triggs (2005), the horizontal and vertical gradients of each frame were calculated, and the magnitude of gradients was determined to be:

$M(x, y) = \sqrt{[G_x(x, y)^2 + G_\gamma(x, y)^2]}$ .....(3)

where G_x and G_γ denote the horizontal and vertical image gradients, respectively.
Gradient orientations were accumulated into spatial histograms, followed by block normalization to produce HOG descriptors:

$h_t = HOG(I_t)$ ..... (4)

PCA was subsequently employed to reduce feature dimensionality and redundancy (Jolliffe, 2002). The compressed HOG representation was obtained using:

$p_t = W_n^T(h_t - \mu)$ ..... (5)

where W_n represents the principal component projection matrix, μ denotes the mean HOG feature vector estimated from training data, and p_t represents the reduced feature vector.
The basic hybrid scheme was kept with 128 components. To explore dimensionality sensitivity, other setups with 32, 64 and 256 components as well as uncompressed HOG representations were tested. The parameters of the PCA were calculated from features of the training set and then used for the validation and testing set.

3.4. Vision Transformer-Based Feature Learning

The second feature extraction branch employed a Vision Transformer to generate contextual visual representations. Following Dosovitskiy et al. (2021), individual frames were divided into non-overlapping image patches, projected into embedding vectors, and combined with positional information. Multi-head self-attention modeled relationships among spatial regions, enabling the extraction of contextual features.
The scaled dot-product attention mechanism was formulated as:

$\text{Attention}(Q, K, V) = \text{SoftMax}[(QK^T)/\sqrt{d_k}]V$ ... (6)

where Q, K, and V represent query, key, and value matrices, respectively, and d_k denotes the dimensionality of key vectors. The resulting frame-level representation was expressed as:

$v_t = ViT(I_t)$ .....(7)

These representations provided contextual visual information complementary to the localized structural characteristics captured by the HOG–PCA branch.

3.5. Hybrid Feature Fusion and Temporal Encoding

The proposed framework integrated HOG–PCA descriptors and ViT embeddings through linear projection and concatenation. The fused representation was defined as:

$f_t = \text{Concat}(W_h p_t, W_v v_t)$ .....(8)

where W_h and W_v denote learnable projection matrices, and Concat represents vector concatenation.
The fused features were projected into a common embedding space and processed through a temporal Transformer encoder:
$Z = \text{TransformerEncoder}(F + P_t)$ .....(9)

where F is the sequence of projected fused features and P_t are the temporal positional embeddings and Z is the encoded visual sequence.

The temporal Transformer captured the sequential dependency crucial for sentence level translation and also modeled relations between consecutive signing movements. Another fusion method, namely the gated fusion, was also explored to find out if adaptive feature integration can be effective in comparison with concatenation-based fusion.

3.6. Sequence-to-Sequence Translation

The temporally encoded visual representations were supplied to an autoregressive Transformer decoder for English sentence generation. Following Vaswani et al. (2017), the decoder employed masked self-attention and cross-attention to generate target tokens conditioned on previously generated words and encoded visual information.
The conditional probability of the target sentence was formulated as:

$$P(Y \mid V) = \prod_{j=1}^{M} P(y_j \mid y_1, \ldots, y_{j-1}, Z) \quad \text{.....(10)}$$

where Y represents the target English sentence, M denotes sentence length, y_j represents the target token at position j, and Z denotes the encoded visual sequence. The empty preceding-token context at the first decoding step is initialized using a start-of-sequence token.
Model optimization employed token-level cross-entropy loss:

$$L_CE = -\Sigma_{j=1}^{M} \log P(y_j \mid y_1, \ldots, y_{j-1}, Z) \quad \text{.....(11)}$$

The loss function encouraged accurate prediction of reference English tokens. During inference, the decoder generated tokens sequentially until an end-of-sequence token was produced or the maximum output length was reached.

3.7. Experimental Evaluation

Translation performance was evaluated using BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and chrF, following Papineni et al. (2002), Lin (2004), and Popović (2015). The proposed architecture was compared with three baseline configurations: HOG–PCA Transformer, ViT Transformer, and parameter-matched ViT Transformer. The parameter-matched baseline was added to investigate if the hybrid architecture can achieve better results than those from model capacity.
For the evaluation of low-resource performance, 10%, 25%, 50%, and 100% of the designated training partition were used. Under these training-data conditions, the proposed hybrid framework was compared with the conventional ViT-only model with BLEU-4.
With respect to architectural ablations, the impact of removing the HOG branch, PCA compression, ViT branch, and temporal Transformer were evaluated. Another experiment was conducted using gated fusion rather than concatenation-based fusion. Configurations with 32, 64, 128 and 256 principal components (PCA) retained were compared with an uncompressed HOG configuration with PCA dimensionality analysis.
The absolute improvement in translation performance was calculated as:

$$\Delta BLEU_4 = BLEU_4(\text{Hybrid}) - BLEU_4(\text{Baseline}) \quad \text{.....(12)}$$

The combined evaluation provided a structured assessment of overall translation accuracy, low-resource performance, architectural contributions, and the influence of feature dimensionality on continuous ISL video-to-text translation.

5. Results and Analysis

5.1 Overall Translation Performance

The translation performance of the HOG–PCA Transformer, ViT Transformer, parameter-matched ViT Transformer, and proposed hybrid HOG–PCA–ViT sequence-to-sequence framework was evaluated using BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and chrF. These metrics assess n-gram precision, sequence overlap, and character-level similarity between generated translations and reference English sentences. The comparative results are presented in Table 1.

Table 1. Comparative translation performance of different architectures

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	chrF
HOG–PCA + Transformer	18.40	11.20	7.10	4.60	23.80	18.20
ViT + Transformer	24.60	16.80	11.70	8.50	30.20	25.40
ViT + Transformer (parameter-matched)	26.10	18.20	13.10	9.70	32.10	27.30
Proposed HOG–PCA + ViT + Seq2Seq	30.80	22.60	17.40	13.20	37.60	32.80

The proposed hybrid framework achieved the best performance on all six-evaluation metrics, with scores of 30.80, 22.60, 17.40, and 13.20 for BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively. The proposed architecture achieved a relative improvement of 55.29% compared to the conventional ViT Transformer, with an absolute improvement of 4.70 points in BLEU-4 score. The same improvement over the ViT Transformer (parameter-matched) was 3.50 points (36.08%).

The proposed hybrid system obtained the best results on all six translation evaluation metrics as shown in Figure 2. The improvement over both the ViT-based baselines suggests that there are opportunities for incorporating handcrafted structural descriptors with the contextual visual representations.

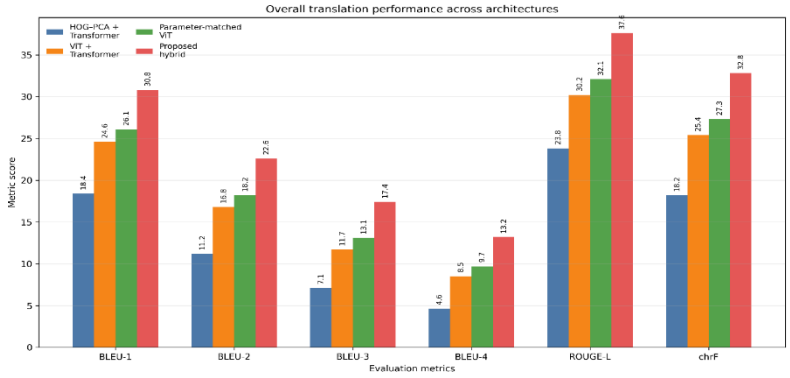


Figure 2. Comparative performance of the HOG–PCA Transformer, ViT Transformer, parameter-matched ViT Transformer, and proposed hybrid HOG–PCA–ViT sequence-to-sequence architecture across BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and chrF metrics.

The proposed model also got ROUGE-L and chrF scores of 37.60 and 32.80 which is 5.50 points better than the parameter matched baseline in both the metrics. Among the different configurations, the standalone ViT configurations achieved higher translation scores on the BLEU-4 metric, with the HOG–PCA Transformer showing the lowest score of 4.60. The results from feature integration indicates that gradient-based structural descriptors and visual representations using the transformer complement each other for sentence-level sign language translation. The comparison to the parameter-matched baseline also suggests that the difference in observed performance is not due to a simple boosting of the number of model parameters, but rather to other architectural factors.

5.2 Low-Resource Translation Performance

The ability of the proposed framework to operate under limited annotated training data was investigated using 10%, 25%, 50%, and 100% of the designated training partition. The BLEU-4 performance of the conventional ViT Transformer and the proposed hybrid architecture is compared in Table 2.

Table 2. Translation performance under different training-data fractions

Training-data fraction	ViT-only BLEU-4	Proposed hybrid BLEU-4	Absolute improvement
10%	2.10	4.30	+2.20
25%	4.20	7.10	+2.90
50%	6.50	10.20	+3.70
100%	8.50	13.20	+4.70

The hybrid architecture achieved higher results in all training-data fractions in comparison to the ViT-only setting. The proposed model also achieved a BLEU-4 score of 4.30 using only 10% of the training data, which outperforms the ViT-only model by 104.76% that scored 2.10 under the same condition. When 25% and 50% of the training-data were available the gains were 2.90 and 3.70 BLEU-4 points, respectively, and the full training partition had the largest absolute difference of 4.70 points.

Figure 3 shows how BLEU-4 is affected by the number of training data. In both architectures, higher translation scores were obtained with higher training fractions; the translation scores of the proposed hybrid were always higher than those of the ViT-only baseline, with the largest absolute improvement of 4.70 points achieved by the proposed hybrid under the full-data training condition.

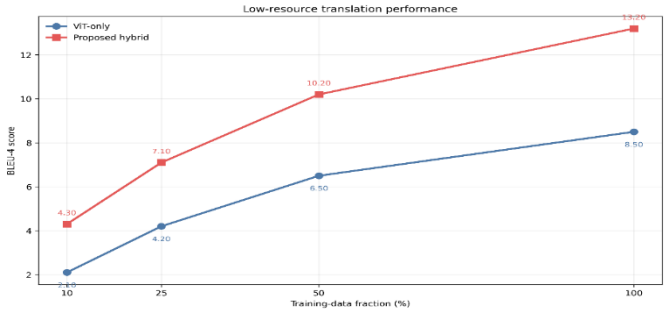


Figure 3. BLEU-4 performance comparison between the ViT-only baseline and proposed hybrid framework across 10%, 25%, 50%, and 100% of the available training data.

In both architectures, they found that the more training data, the better. However, the hybrid model consistently outperformed, even in the most stringent training setting. The relative improvement at 10% training-data availability seems to indicate that the combination of HOG–PCA descriptors might be very useful in cases where task-specific supervision is scarce. Nevertheless, the absolute BLEU-4 scores at the smaller training fractions also suggest that it is still quite difficult to generate accurate sentences in very limited training conditions.

5.3 Ablation Analysis of Architectural Components

An ablation analysis was performed to determine the effects of the individual architectural components on the translation performance. Component removal was applied to the HOG branch, PCA compression, ViT branch and temporal Transformer. In addition, the gated feature fusion was contrasted with the concatenation-based feature fusion used in the complete architecture. Results are given in Table 3.

Table 3. Ablation analysis of the proposed architecture

Model configuration	BLEU-4	ROUGE-L
Complete HOG–PCA + ViT + Seq2Seq	13.20	37.60
Without HOG branch	9.70	32.10
Without PCA	11.40	35.20
Without ViT branch	4.60	23.80
Without temporal Transformer	6.80	27.40
Gated fusion instead of concatenation	12.60	36.90

The complete hybrid architecture obtained BLEU-4 13.20 and ROUGE-L 37.60 results, which were the best scores among all the variants evaluated. The removal of the ViT branch led to the most significant drop in BLEU-4, which dropped by 8.60 points and ROUGE-L, which dropped by 13.80 points. This outcome reflects the significance of using visual representations obtained from the transformer in the proposed architecture.

The architectural ablation results shown in Figure 4 are used to show the relative contribution of the components of the proposed framework. The BLEU-4 and ROUGE-L scores decreased the most when the ViT branch was removed, followed by the removal of the temporal Transformer. The full architecture had the best performance, which benefited from structural feature extraction, contextual visual representation, and temporal modeling.

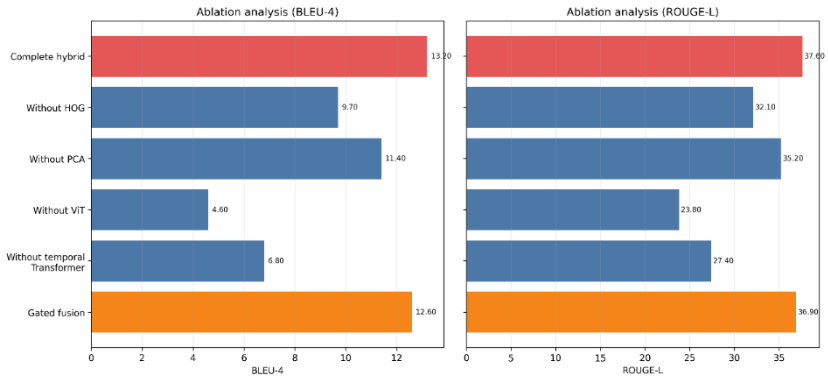


Figure 4. Effect of architectural component removal and fusion-strategy modification on BLEU-4 and ROUGE-L performance of the proposed HOG–PCA–ViT framework.

Removing the temporal Transformer resulted in a drop in BLEU-4 by 6.40 and a drop in ROUGE-L by 10.20, giving a score of 6.80 and 27.40 respectively. These differences illustrate the effect of temporal contextual modeling in the translation of continuous sequences of signs. Removing the HOG branch resulted in a drop of 3.50 points of BLEU-4, while removing PCA compression gave a smaller drop of 1.80 points, which indicated that the complementary gradient information and dimensionality reduction was beneficial.

The results for gated fusion were BLEU-4 of 12.60 and ROUGE-L of 36.90. This setup was still competitive, but concatenation still gave 0.60 BLEU-4 and 0.70 ROUGE-L improvements. The ablation results overall show that the best results were achieved when combining the handcrafted spatial features, deep visual features, temporal contextual modeling, and concatenation-based feature fusion.

5.4 Effect of PCA Dimensionality on Translation Performance

The effects of PCA dimensionality on the translation performance was investigated using 32, 64, 128 and 256 principal components. Another configuration without PCA was added to evaluate the impact of not compressing the HOG representation. The BLEU-4 scores are shown in Table 4.

Table 4. Effect of PCA dimensionality on translation performance

PCA configuration	BLEU-4
32 components	9.80
64 components	11.60
128 components	13.20
256 components	12.80
Without PCA	11.40

Translation performance was improved with the PCA dimension up to 128 components, and the BLEU-4 scores were 9.80, 11.60, and 13.20 for 32, 64, and 128 components, respectively. When the dimensionality was extended to 256 components, the score dropped slightly to 12.80 indicating that further principal components did not give any additional translation gain.

As displayed in Figure 5, BLEU-4 score improved with the number of principal components kept increased from 32 to 128, with the best score of 13.20 at 128 principal components. A higher component number (256) led to a slight decrease in performance and the uncompressed HOG configuration gave a lower score of 11.40. These findings revel 128 components as the best choice from the ones tested.

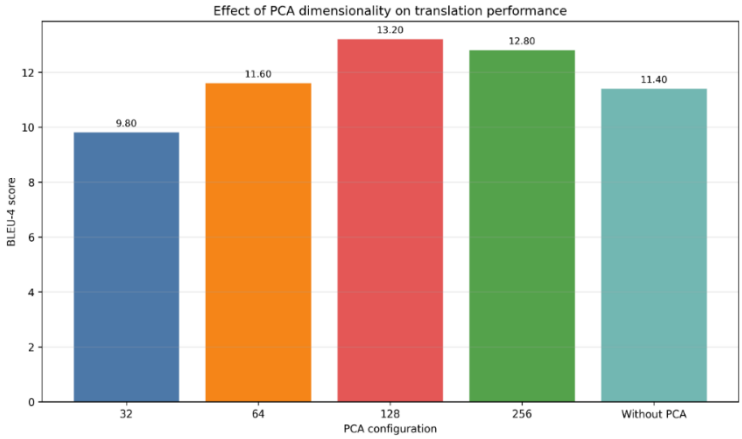


Figure 5. Effect of retaining 32, 64, 128, and 256 principal components, compared with uncompressed HOG descriptors, on BLEU-4 translation performance.

With the uncompressed HOG configuration, BLEU-4 score was 11.40, 1.80 lower than with the 128-component configuration. This discrepancy seems to indicate that PCA-based compression could be used to influence the performance of a translation system by compressing visual information that has been derived from a gradient. The lower scores achieved with 32 and 64 components reflect a possible loss of information with more aggressive dimensionality reduction, while the slight drop in score at 256 components indicates that the benefit of including more components is limited.

The configuration with 128 principal components yielded the best translation score out of the range of configurations tested and this dimensionality was determined to be the optimal size of the HOG feature branch. This outcome provides evidence for the application of an intermediate feature representation which combines dimensionality reduction with retaining translation-relevant spatial information.

6. DISCUSSION

Results show the benefits of using handcrafted HOG–PCA descriptors and Vision Transformer (ViT) embeddings in the sentence-level Indian Sign Language (ISL) video-to-text translation. The proposed hybrid architecture obtained a BLEU-4 score of 13.20 which is 4.70 and 3.50 points higher than the BLEU-4 score of conventional ViT Transformer and parameter matched ViT Transformer, respectively. Further improvements have been shown in ROUGE-L (37.60) and chrF (32.80), which show an improvement in the correspondence between the generated and reference sentences. The findings indicate that the gradient information at the local level can be used to supplement the visual representations in the context, thus offering extra discriminative features for continuous gesture interpretation.

The experiments with low resources showed a similar improvement for all the fractions of training data. The hybrid model outperformed the ViT-only architecture by 104.76% at 10% training availability with BLEU-4 score of 4.30 when compared with 2.10 of the ViT-only architecture. However, in the case of absolute translation accuracy, under highly restricted supervision, the accuracy was still limited. Ablation results also confirmed the role of the different architectural components. ViT removal led to BLEU-4 of 4.60 and the removal of temporal encoding led to BLEU-4 of 6.80. The HOG branch was used for the other performance and it was found that 128 components gave the best performance in PCA dimensionality analysis. All these results suggest the complementary nature of spatial representation, feature compression and temporal contextual modeling.

The results are in line with the previous studies which highlighted attention-based architecture for sign language translation in continuous mode. Camgoz et al. (2020) have shown the effectiveness of Transformer-based sequence modeling and Zhou et al. (2023) showed the benefits of visual–language pretraining for gloss-free translation. Likewise,

Yin et al. (2023) pointed out that it is necessary to pay attention to context when interpreting continuous signing sequences. We follow these research directions here by adding handwritten gradient descriptors together with learned visual embeddings prior to the temporal encoding.

As for low-resource translation, Joshi et al. (2025) showed that linguistically informed pose-sequence learning can be done with PoseStitch-SLT. The present framework explores spatial feature integration in contrast to their approach that focuses on pose based structure supervision. The results also agree with the significance of controlled architectural comparison that was highlighted by Sincan et al. (2025). But you need to have comparable data, data preprocessing and evaluation protocols in order to compare numbers directly with these studies.

The basic theoretical significance is that handcrafted and deep visual representations are able to deliver complementary information in a single translation architecture. The results of removing either HOG, ViT or temporal encoding indicate that the combination of localized structural features and contextual spatial and sequential representation should be used to achieve effective sentence-level translation. Moreover, the PCA analysis shows that the higher dimensions of the features don't necessarily give better translation results.

From a practical perspective, the framework can serve as a foundation for constructing ISL translation systems as there is only a small amount of annotated data available. It also possesses the modular style that indicates various feature branches or temporal components can be changed independently. This flexibility will assist future research in the systems for efficient translations for education, accessibility and human computer interaction. There are still some problems that need to be tested in terms of reliability of translation, latency and robustness before actual deployment.

Some caveats are in order. First, automatic lexical similarity measures cannot be used to completely determine the semantic correctness, grammatical adequacy, and nonmanual aspects of language preservation. Second, the relatively low BLEU-4 scores obtained using limited training data suggest that there are still issues when trying to create accurate sentences given limited supervision. Third, the reported comparisons cannot give confidence intervals and there is no information on variations of performance between repeated runs. In addition, the evaluation provided does not verify generalization on independent data, across unseen signers, recording setups or in real world communication situations. The efficiency of computation and inferences are also poorly understood.

Future investigations are needed to test the proposed framework with independent ISL data and signer-disjoint protocols to test for generalization. Using hand landmarks, facial expressions and bodypose information might help to better capture linguistic information that is not sufficiently captured by the frame-level visual embeddings. Quality of translations may be further enhanced under limited supervision by using contrastive pretraining, self-supervised representation learning and language-model-assisted decoding. Further study is needed on the statistical significance, semantic evaluation, human evaluation based on fluent ISL users and computational optimization. Practical deployment and tests for robustness under various signing conditions would also help to determine the suitability of the framework for the use of accessible communication technologies.

7. CONCLUSION

The present work proposes and investigates a hybrid Histogram of Oriented Gradients–Principal Component Analysis (HOG–PCA) and Vision Transformer (ViT) based sequence-to-sequence (S2S) approach for video-to-text translation of low resource Indian Sign Language (ISL) videos. The design was intertwined with handcrafted spatial descriptors and contextual visual embeddings which were then temporally encoded using Transformers and autoregressively translated to English sentences. Comparative model evaluation, low resource experiments, architectural ablation study and PCA dimensionality analysis were systemically performed on the framework.

The experimental results indicated that the proposed hybrid framework outperformed other frameworks in terms of the translation performance under all evaluated metrics, achieving the BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores of 30.80, 22.60, 17.40 and 13.20, respectively, as well as ROUGE-L and chrF scores of 37.60 and 32.80. The BLEU-4 gain of +4.70 points for this combination of complementary visual representations compared to a conventional ViT Transformer and +3.50 points compared to a parameter-matched configuration further highlighted the performance benefits. The superiority of the performance was also validated by low-resource experiments by consistently showing better performance at varying training-data fractions. Ablation analysis showed that the use of ViT representations, temporal modeling and handcrafted features is important, with the best PCA dimensionality evaluation being 128 principal components.

The main novelty of this work is the proposed method, which is based on the integration of small handcrafted features and rich context visual features for continuous ISL translation. However, the quality of the translations is still limited, and more testing is needed to determine the extent to which it can be generalized and used reliably in practice. Future research will focus on multimodal visual features, self-supervised pretraining, Signer independent evaluation, Semantic translation assessment, and Computational optimization. In summary, the results constitute a stepping stone towards the creation of better and data-saving ISL video-to-text translation systems.

REFERENCES

1. Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10023–10033). https://openaccess.thecvf.com/content_CVPR_2020/html/Camgoz_Sign_Language_Transformers_Joint_End-to-End_Sign_Language_Recognition_and_Translation_CVPR_2020_paper.html

2. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 1, pp. 886–893). IEEE. https://doi.org/10.1109/CVPR.2005.177
3. De Coster, M., Van Herreweghe, M., & Dambre, J. (2020). Sign language recognition with transformer networks. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 6018–6024). European Language Resources Association. https://aclanthology.org/2020.lrec-1.737/
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy
5. Gong, J., Foo, L. G., He, Y., Rahmani, H., & Liu, J. (2024). LLMs are good sign language translators. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18362–18372). https://doi.org/10.1109/CVPR52733.2024.01738
6. Hamidullah, Y., van Genabith, J., & España-Bonet, C. (2024). Sign language translation with sentence embedding supervision. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 425–434). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.40
7. Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer. https://doi.org/10.1007/b98835
8. Joshi, A., Agrawal, S., & Modi, A. (2023). ISLTranslate: Dataset for translating Indian Sign Language. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 10466–10475). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.665
9. Joshi, A., Mohanty, R., Kanakanti, M., Mangla, A., Choudhary, S., Barbate, M., & Modi, A. (2024). iSign: A benchmark for Indian Sign Language processing. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 10827–10844). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.643
10. Joshi, A., Sharma, V., Singh, S., & Modi, A. (2025). PoseStitch-SLT: Linguistically inspired pose-stitching for end-to-end sign language translation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 13834–13853). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.698
11. Kim, J., Jeon, H., Bae, J., & Kim, H. Y. (2025). Leveraging the power of MLLMs for gloss-free sign language translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21048–21058). https://openaccess.thecvf.com/content/ICCV2025/html/Kim_Leveraging_the_Power_of_MLLMs_for_Gloss-Free_Sign_Language_Translation_ICCV_2025_paper.html
12. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013/
13. Low, J. H., Sincan, O. M., & Bowden, R. (2025). SAGE: Segment-aware gloss-free encoding for token-efficient sign language translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 5011–5020). https://openaccess.thecvf.com/content/ICCV2025W/MSLR/html/Low_SAGE_Segment-Aware_Gloss-Free_Encoding_for_Token-Efficient_Sign_Language_Translation_ICCVW_2025_paper.html
14. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135
15. Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049
16. Shahin, N., & Ismail, L. (2024). From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: Survey, taxonomy and performance evaluation. Artificial Intelligence Review, 57(10), Article 271. https://doi.org/10.1007/s10462-024-10895-z
17. Sincan, O. M., Low, J. H., Asasi, S., & Bowden, R. (2025). Gloss-free sign language translation: An unbiased evaluation of progress in the field. Computer Vision and Image Understanding, 261, Article 104498. https://doi.org/10.1016/j.cviu.2025.104498
18. Tanzer, G., Shengelia, M., Harrenstien, K., & Uthus, D. (2024). Reconsidering sentence-level sign language translation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 6262–6287). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.360
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
20. Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., & Zhao, Z. (2023). Gloss attention for gloss-free sign language translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2551–2562). https://doi.org/10.1109/CVPR52729.2023.00251
21. Zhou, B., Chen, Z., Clapés, A., Wan, J., Liang, Y., Escalera, S., Lei, Z., & Zhang, D. (2023). Gloss-free sign language translation: Improving from visual-language pretraining. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20871–20881). https://doi.org/10.1109/ICCV51070.2023.01908