© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719448

Temporal And Multimodal Enhancement of Semantically Interpretable Attention for Online Hand Gesture Recognition

SANABOINA CHANDRA SEKHAR[1], SAIBA TEJA SRI[2]
[1]*Assistant Professor, Department of CSE, UCEK(A) JNTU KAKINADA*
[2]*PG Scholar, Department of CSE, UCEK(A) JNTU KAKINADA*

Abstract- ***Hand gesture recognition plays significant role in HCI applications and also it has been incorporated in the fields of Virtual Reality (VR), Augmented Reality (AR) and Robotics. Gestures in the real world are continuous and under these conditions, any recognition online would be very difficult since there are two types of gesture classes to be recognized as well as the temporal limits of these classes of gesture. Current state-of-the-art deep learning based on sliding windows and attention mechanisms yield piece-wise and "jagged" predictions resulting in more false alarms as well as incorrect gesture boundaries. Single-modality methods are not beneficial for recognition purposes because they do not return full-fledged gestures of the gesture. It is a reproduction of a cross-attention-based online hand gesture recognition model based on Joint Collection Distance (JCD) and Frame Vector (FV) that reproduces attention-based models. Taking the temporal refinement approach is proposed to reduce the noise and enhance the detection of the boundaries. To extend this into multimodal, visual appearance features of video frames are added in through an existing CNN which can learn structural and appearance features. Experimental findings of IPN Hand Gesture Dataset indicate that the Temporal refinement enhanced the Detection Rate of 0.9184 to 0.9288, False Positives of 39 to 14, and Mean IoU of 0.7375 to 0.8153 which showed that the gesture recognition and temporal localization were more accurate.***

Index Terms- ***Hand Gesture Recognition, Multimodal Learning, Attention Mechanism, Temporal Refinement, Joint Collection Distance (JCD), Frame Vector (FV), Deep Learning, Sliding Window.***

I. INTRODUCTION

Human-Computer Interaction (HCI) has become the focus of the present smart systems, and hand-gesture has the natural and touchless properties, and investigated from the many hands interact modes. Gesture-based interfaces have now been observed in various areas such as virtual reality (VR) and augmented reality (AR), robotics, video games, intelligent environments and sign language interpretation. The more these applications are used, the more real-time recognition systems are needed that are accurate.

This challenge is defined by one key difference: pre-segmented gestures can be classified offline whereas online systems have to work with a sequence of raw gestures as they are made, and, in parallel, determine which gesture is made and when, which is much more difficult and rendered problematic by transitions between gestures, timing uncertainty, background movement, and real-time constraint.

First Systems utilized custom characteristics and conventional machine learning, working fairly well in a managed setting, but incapable of dealing with real-world changes and variations. The spatiotemporal model in [29] was helpful but was not strong enough in a variety of circumstances.

The revolution of deep learning paved the way for direct learning of spatial and temporal representations in the form of convolutional, recurrent and temporal convolutional architectures and brought significant improvements in performance and generalization [12,7,27].

Skeleton based techniques turned attention on the other side and started to use the changes of the image joints for recognition rather than using the raw pixels and paying attention to illumination and cluttered context. Graph based methods like two-stream adaptive graph convolutional networks [17] and

directed graphs neural networks [18] were able to learn to jointly represent joint-level spatial structure and temporal motion evolution.

More recently, focus has been on attention mechanisms, which offer more interpretable model behaviour and greater representations. Although transformer structures have been shown to be able to model long-range temporal dependence [14], more focus on the most informative spatial and temporal cues is possible by using attention-enhanced graph networks [16] or adaptive graph-based recurrent models [17].

This is directly related to the semantically interpretable attention model described in [1] which uses two hand landmark features – Joint Coordinate Distance (JCD), and Frame Vector (FV) to enable real-time gesture recognition. Despite its excellent performance, it still has some pitfalls to avoid: prediction limitation, prediction accuracy issues and low utilization of visual appearance information still lead to inaccuracies in the real deployment area.

The paper fills this gap with a Temporal and Multimodal Enhancement system that is based on the system of [1]. A temporal refinement module is presented as a post-processing step to remove short sporadic predictions, bring together misplaced parts and process the temporal boundaries to more coherent outputs.

It also includes a multimodal dual-attention fusion mechanism, which is based on skeleton motion features and visual appearance features (obtained by a pre-trained ResNet-18 network) and jointly uses structural hand motion and appearance-level context. The framework increases Detection Rate (0.9184 to 0.9288), minimizes False Positives (39 to 14) and increases Mean IoU (0.7375 to 0.8153) on the IPN Hand Gesture Dataset [2], proves the effectiveness of the proposed improvements to online gesture recognition.

The rest of this paper is structured in the following way. Section II explores the existing literature concerning gesture recognition, temporal modelling and multimodal learning. In Section III, detailed methodology, comprising feature extraction, feature fusion and feature temporal refinement, is explained. Results and analysis of the experiments are found in Section IV. Directions for the future and summary of the paper are given in Section V.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

A lot of progress has been made in the past years concerning the recognition of hand gestures, starting from handcrafted rule-based pipeline, to more advanced architectures such as temporal modelling, attention model and multimodal sensing. It is this part which builds on that development and places the proposed work within the current context.

A. Introduction to early foundations and temporal modeling.

In the early research on continuous gesture recognition, it was found that these problems cannot be solved separately since they are closely related. In the case of [29] this was the shared spatiotemporal framework, looked upon in the same pipeline, which set down important conceptual bases but was not highly scalable due to being based on handcrafted representations.

To address this, researchers have come up with more expressive temporal modeling schemes: hierarchical recurrent neural networks [19] proved that multi-level temporal abstraction enhances skeleton-based gesture recognition, short-term sampling networks [13] provided that targeted window sampling is more discriminative motion cues, and temporal convolutional networks [7] achieved high sequential modeling with dilated convolutions, not requiring the computational scalability of recurrence.

B. Deep Convolutional and Recurrent Structures.

Among the highlights was the adoption of a 3D-CNN that learned spatial and temporal properties of videos jointly in one hierarchy of learning, outperforming the former two-stream architectures that learned properties of appearance and motion separately [27]. Frequent 3D CNN architecture [4] extended it to the real-time online recognition which can be used practically in an interactive environment.

The addition of recurrent connections directly in the convolutional structures [8] introduced a more

compact approach and faced the sequential learning of features without the need of recurrent modules. However, they are very simple to decorrelate in the temporal dimension and are often sensitive to temporal long-range correlations in long sequences of gestures, making the methods sometimes very computationally intensive.

Part III: Skeleton Based and Graph Structured Recognition.

Skeleton-based techniques are gradually gaining in importance in gesture recognition as the principle by which features are extracted from the gestures is based on the geometric information of the joints and motion trajectory, and does not rely on the raw data of image's pixels, so they are more immune to the changes of illumination and background noise.

Two-stream adaptive graph convolutional networks [17] proposed learnable graph topologies, which are more flexible than the fixed counterparts related to the spatial joint dependencies and temporal dynamics. In addition, spatial modeling was further enhanced with the use of directed graph neural networks [18] which modelled asymmetric joint relationships with directed graph structures.

Attention-enhanced graph convolutional LSTM networks [16] that were based on graph-based spatial modelling, and recurrent temporal processing and attention-guided selection built additional recognition strength. Taking one step further, it treats both spatial and temporal dependencies together in a common attention context, as in the spatial-temporal synchronous transformer [14] showing that transformers can be used successfully in the context of skeleton-based gesture data.

Doesn't this feature make you think of Deep Learning and Attention Mechanisms and Interpretable Representations?

The attention mechanisms have been more quickly incorporated in gesture recognition since they had been more generally successful in sequence modeling. Previously, it has been demonstrated that, in the field of skeleton-based recognition to build up a more powerful and faster basis of skeleton-based recognition, the perfectly designed attention module can also result in higher correctness at the cost of lower computation load [23].

Enriched graph convolutional networks (GCNs) [22] showed that graph convolutional layers (GCLs)' pattern of activation is important for creating generalizable and recognizable features.

The one which is closest to it is the semantically interpretable attention model of [1] which combines JCD and FV features of MediaPipe hand landmarks with cross-attention modules yielding competitive performance with informative information about the spatial and temporal features which guide recognition decisions.

These strengths can be expressed, but dis-grained predictions, imprecise time limits and unexploited “the way things look” information is the scourge that never escapes this well-designed yet unsound system.

C.Temporality of stability in the processes of refining and prediction.

Great attention was paid to the consistency of gesture predictions through time because in practice it is very easy to get to a point where gestures are predicted inconsistently, whether this is by being present at a certain time or absent at other ones.

Temporal hierarchical dictionary-guided decoding [5] had a hierarchical decoding step but sequentially refined the segment boundaries, and could also support finer localization of the result compared with flat method.

In order to process gesture sequences, deep models and temporal segment neural networks [15] segmented the gesture sequences into smaller, and then multiplied and summed the results.

An integrated 3D CNN convolutional LSTM framework [6] was used to capture the local spatially-temporal dynamics and longer-range sequence dynamics in one design. To alleviate the weight of the 3D convolutions, gate-shift networks [26] proposed the learnable temporal gating on the 2D convoluted feature maps.

Despite such efforts, there are still a lot of false detections and nuanced differentiation between gesture and background motion, particularly in unconstrained settings where it is difficult, if not impossible, to successfully and neatly differentiate between them.

D. Multi modal learning of Gestures.
Owing to their complementary nature, the fusion of the two types of sensing modalities is gaining more and more evidence as the optimal approach to overcome the drawbacks of one of the sensors.

Multimodal gesture recognition on large scale using heterogenous networks [9] showed that with the joint use of skeleton data, RGB and depth data, there was always more consistency and high performance with respect to different gesture types.

In fact, we observed that although only skeleton features are used at inference time, the use of visual appearance during training improves the learned representations; thus, multimodal training can be beneficial for multimodal inference.

The need for spatiotemporal feature learning with 3D convolutions [27] and two-stream recurrent nets' ability [20] of demonstrating that 2 streams of complementary features (visual and motion) are greater than single-pathway video processing also demonstrated the importance of handling both spatial and temporal streams of video. All these results help to demonstrate the need for a multimodal system to combine the skeletons and vision modes.

E. Benchmarks, surveys and continuous challenges.
There has been a major role played by benchmark datasets and systematic reviews in research priority formation. The central benchmark used for this work is the multi-gesture IPN Hand dataset [2] that provides a continuous sequence of multi-gesture data with frame-wise temporal information which can be used for online recognition evaluation.

With the publication of large-scale datasets like NTU RGB+D 120 [24] and NTU RGB+D [25] the evaluation processes for skeleton-based systems have been established. Even when trained with partial supervision, effective temporal models can be built by Continuous sign language recognition trained iterative [28] and jointly learning of spatio-temporal graph operations trained by unified graph convolution optimization [21] showed that more spatio-temporal relationship is learned more effectively than sequential learning.

Since 2018, systematic review of hand gesture recognition [30] and sensory glove-based sign language systems [31] list down common factors such as real-time processing, temporal segmentation error, руки variability and single-modality as the most prevalent open issues within the boundaries of the field. Work to using lightweight CNNs to produce real-time detection [11] has shown that it is possible to achieve a balance between speed and accuracy, but it is difficult to be consistent with different styles of gestures.

MultiStream 1D CNN fusion [12] was proven to be more robust as parallel processing of the streams was used, however is not quite trivial to balance the streams during fusion. In each of these directions the literature prenatal up the same gaps in the continuous and exact temporal boundary’s location, the effective combination of multimodal, which motivated the framework developed in this paper.

III. WRITE DOWN YOUR STUDIES AND FINDINGS

A.System Overview
The research article introduces a time- and multiple-modal attention-based hand gesture recognition improvement framework which boosts the ability to semantically interpret attention-based hand gesture recognition in online continuous conditions. The framework is based on four key parts, including skeleton-based motion features, RGB visual appearance features, a cross-attention fusion system and a temporal polishing system, which are structured into a sequential modular process that provides stable and strong recognition with real-time constraints.

It is possible to mention several design options in advance. ResNet-18 is trained without fine-tuning and applied in its pretrained format, as it is a fixed feature provider only. MediaPipe Hands is a purpose-

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719448

built preprocessor of landmark, and all skeletal features are computed by deterministic mathematical integrations instead of learned transformations, making this portion of the pipeline completely transparent, and reproducible.

The process works in a logical sequence of steps - starting with the acquisition of raw video frames and landmarks, computing the features and footing up the multimodal fusion, followed by temporal refinement and ultimate generation of gesture output - so that the information in both the modalities is always fused. The general construction of the suggested system is depicted on Figure. 1.

Fig 1: Proposed Multimodal Online Hand Gesture Recognition Architecture

B. Dataset and Preprocessing Pipeline

The suggested framework develops its multimodal data based on the IPN Hand Gesture Dataset, a widely used standard in continuous recognition of gestures on-line. The dataset represents a video sequence of RGB videos of various gesture types that have been annotated with label of gesture classes at a frame-level and the exact start and end frame number.

This raw data is then converted into training and evaluation-ready inputs in a structured multi-stage preprocessing pipeline, including video frame extraction, landmark detection, skeleton feature computation, visual feature extraction, and sliding window generation - so that there is no change in processing or a temporal discrepancy between both modalities.OpenCV frames are processed to extract them and then the frames are resized to 224 x 224 pixels to restore uniform spatial resolution.

MediaPipe Hands then identifies 21 joint coordinates in 3D, per frame, that comprise the skeletal backbone of computing features. These landmarks are the basis of two complementary representations, Joint Coordinate Distance (JCD) features, which project all pairwise geometric distances among all hand joints onto a 210-dimensional static spatial representation, and Frame Velocity (FV) features, which project inter-frame landmark displacement onto a 63-dimensional motion representation. Combining the two results in a combined 273-dimensional skeleton feature vector at each frame.

At the same time, a pretrained ResNet-18 with its classification head removed predicts a 512-dimensional visual appearance vector of each frame, which captures high-level attributes of the hand and its context. Both sets of feature sequences are then subjected to sliding windows of 16 frames with a stride of 1 to produce the temporally structured input samples.

With current matching of frame to ground-truth annotations each window is labelled with its functionality (enough overlap to match) to give it the corresponding gesture type and all larger windows are labelled with a background tag. The whole process of preprocessing and sliding window generation pipeline is explained in Figure. 2

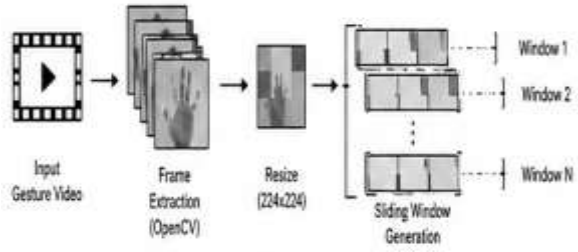

Fig 2: Video Preprocessing and Sliding Window Generation Pipeline

C. Video Input and Skeleton Feature Extraction

Video input is constantly grabbed and divided into sequential frames which are resized and normalized in order to be similar in the pipeline. Each frame undergoes MediaPipe Hands to identify and track 21 hand landmarks, expressed in three-dimensional coordinates, across joints of the fingers, fingertips, the palm and the wrist.

These landmarks yield two complementary types of skeleton features, which address two aspects of hand motion.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719448

JCD features capture the geometric nature of the hand that is not dynamic by calculating the Euclidean distance between each pair of joints in a frame as expressed in Equation. (1):

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \qquad (1)$$

where d_ijthe Euclidean distance between landmark points iand j, and

(xⓜ,yⓜ,z)denote their three-dimensional coordinates. This generates a 210-dimensional spatial representation per frame coding the instantaneous hand shape.

The FV features shift the attention to motion dynamics by quantifying the relative motion of each landmark across successive frames, as given in Equation. (2):

$$FV_t = L_t - L_{t-1} \qquad (2)$$

where L_tand L_(t-1) are the coordinate vectors of the landmarks at the consecutive time steps tand t-1. The result is a 63-dimensional motion descript that expresses hand movement speed and direction over time.

The two types of features are combined into a single 273-dimensional feature of each frame, as illustrated in Figure. 3, integrating the approach to a skeletal gesture representation of JCD with the FV approach of motion dynamics over time into a more comprehensive representation.

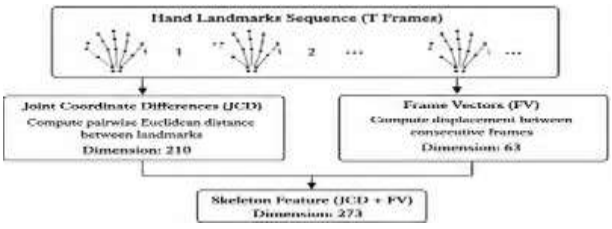

Fig 3: Skeleton Feature Extraction Process

D. Visual Feature Extraction using ResNet-18

Although geometrically expressive, skeleton features will never be able to reproduce the rich visual detail of raw video frames. To resolve this, the proposed framework will have a specific visual feature extraction branch, which works with RGB frames directly, applying the pretrained ResNet-18 deep convolutional neural network.

In each frame, which is uniformly decomposed into 224 x 224 pixels each, every frame is fed through the sequential processing stages of ResNet-18, starting with an initial convolution and max-pooling stage that detects the low-level features, followed by four stages of residual processing that gradually improve the abstractions and semantically significant representations, and a final global average pooling stage that summarizes the spatial information of a frame into a concise representation. The last classification layer is removed to have the network acting simply as a feature extractor.

The characteristic feature of ResNet-18 is its residual connections that form shortcut connections that maintain gradient flow along deep layers and enable the network to simultaneously capture fine-grained low-level features as well as high-level semantic patterns without degradation of semantic representation.

degradation. After global average pooling, the network generates a 512-dimensional feature-vector every frame that encodes visual features like that of hand shape, surface texture and contextual appearance indicators that cannot be encoded by skeletal landmarks alone.

This visual stream is the second input to the multimodal fusion module, which is more complementary to the structural motion data provided by the skeleton branch with the appearance level information provided by the video content. The visual feature extracting model based on ResNet-18 is depicted in the figure below. Figure. 4.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719448

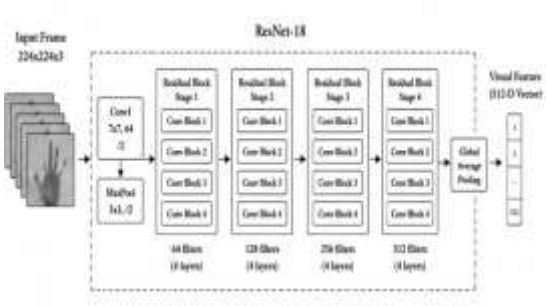

Fig 4: Visual Feature Extraction using ResNet-18

E.Cross-Attention-Based Multimodal Fusion

Instead of simply doing concatenation of the two feature streams, the proposed framework does the integration between the two using a multimodal dual-attention fusion mechanism which is designed to actively learn meaningful cross-modal relationships.

At the start, the projections of the two streams are into a common projection space: a fully connected layer maps from the 273-dimensional skeleton features to a 128-dimensional hidden space, and a fully connected layer maps from the 512-dimensional visual features to the same 128-dimensional hidden space, both modalities thus operating at the same scale without any interaction.

For each of the projected streams, temporal attention modules are then trained to decide which of the time steps is most informative in each modality, thus also capturing motion dependencies throughout the gesture.

After this independent refinement, a cross-modal attention mechanism is used to do directed information exchange between the two streams, by the scaled dot-product formulation as defined in Eq. (3).

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

The query, key and value matrices are denoted by Q, K, and V respectively, and dk is the dimensionality of the key used for scaling purposes.

In this configuration the skeleton features are used as a query and the visual features are used as a key / value representation. This is encapsulated in the structure and motion cost that are stored in the skeleton stream, which explicitly influences which visual data to activate, and thus explicitly influences what is emphasized at each time and what is not.

This ultimately integrated vector combines the structural hand configuration with information about the motion dynamics over time and appearance level contextual information, that is sent to the classification module. The overall cross-attention based multimodal fusion process is illustrated in Figure. 5.

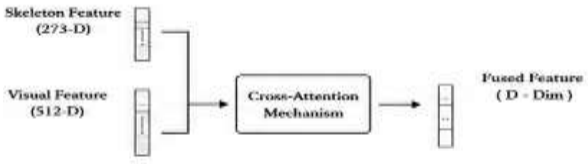

Fig 5: Cross-Attention-Based Multimodal Fusion

F. Gesture Classification and Temporal Inference

The fused multimodal feature vector is passed into a fully connected classification layer that maps it onto probability scores across all gesture classes. The predicted label for each input window is determined by selecting the class with the highest probability, as expressed in Eq. (4):

$$\hat{y} = \arg\max P(y \mid x) \qquad (4)$$

where y^denotes the predicted gesture class, cindexes over all available gesture categories, and P(y=c|x)is the model's estimated probability that input xbelongs to class c.Since the model processes continuously generated overlapping temporal windows rather than isolated clips, each window receives both a gesture label and a corresponding pair of temporal boundary positions — start and end frame indices — enabling uninterrupted real-time recognition without requiring explicit input pre-segmentation.

However, this window-level inference introduces practical limitations. Raw outputs at this stage are inherently imperfect — inconsistent labels, brief spurious detections, and misaligned segment boundaries are common occurrences that, left

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719448

uncorrected, accumulate into fragmented and unreliable recognition outputs.

To address this, the raw per-window predictions are passed to the temporal refinement module, which applies a structured set of post-processing operations to improve coherence, stability, and boundary precision across the final recognition output, as shown in Figure. 6.

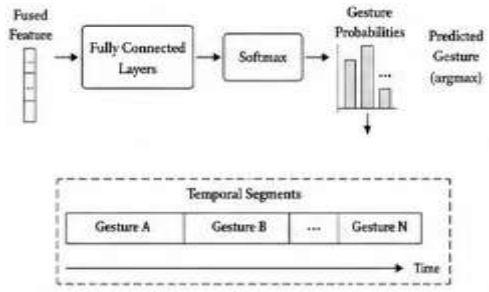

Fig 6: Gesture Classification and Inference Process

G. Temporal Refinement and Noise Reduction

Raw frame-level predictions of continuous gesture streams are naturally noisy and inconsistent even with a trained model. To tackle this, the suggested structure will include a temporal refinement module, which functions as a post-processing step to directly apply to the prediction outputs, without altering the underlying model, via three sequential operations.

The noise segment removal operation eliminates predicted segments with a length that is less than a set minimum length. These brief segments are usually the result of transient landmark detection errors, or temporary background motions that shallowly mimic real gestures and have no actual instance of gesture.

Segment merging, the second operation, determines adjacent predictions that are of the same gesture class and separated by a only.

short temporal gap - a fragmentation pattern that is often found when overlapping windows are used to process a single continuous gesture in a non-uniform manner. These adjacent fragments are united into a single continuous forecast, creating a more temporally coherent output.

The third step, boundary refinement, increases the start and end points of each of the predicted segments slightly to augment temporal overlap between the gesture intervals in the ground truth and those in the predicted gesture. This enhances the localization accuracy without compromising on the class labels assigned.

Combined in this order, these three operations are what turn noisy and fragmented raw predictions to smoother, more stable, and bounded gesture predictions, unambiguously decreasing false positive detections and enhancing temporal localization of continuous recognition sequences. Figure. 7 shows the entire process of temporal refinement.

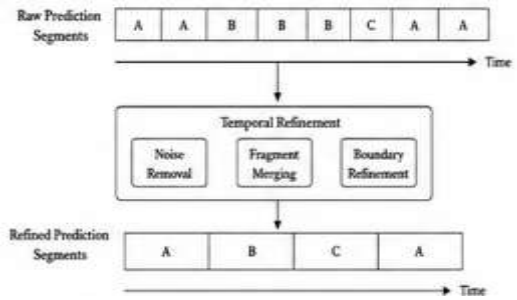

Fig 7: Temporal Refinement Process

H. Evaluation and Performance Analysis

The framework is tested on the next IPN Hand Gesture Dataset on the continuous online recognition conditions which are close to the real-life implementation. Predictions in gesture are obtained at each test sequence at the raw inference step, as well as at the stage of temporal refinement thereof, then compared to frame-level ground-truth annotations using three common measures.

The Detection Rate (DR) is the ratio of the number of true gestures detected by the system to the number of ground-truth gestures (Eq. (5):

$$DR = \frac{Correctly\ Detected\ Gestures}{Total\ Actual\ Gestures} \quad (5)$$

Mean Intersection over Union (Mean IoU) evaluates temporal boundary quality by measuring the overlap between predicted and ground-truth gesture segments, as defined in Eq. (6):

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719448

$$IoU = \frac{Intersection}{Union} \quad (6)$$

where P_iand G_idenote the predicted and ground-truth intervals for the i-th gesture instance and Nis the total number of instances evaluated.

False Positives (FP) the number of predictions that are not associated with any real gesture a direct indication of how the system is prone to give spurious predictions. The combination of the increased DR and Mean IoU values can be regarded as a sign of superior recognition performance and more accurate localization, whereas the lower number of FP can be interpreted as a cleaner and more consistent output stream.

In order to appropriately separate the influence of each component, the evaluation is done individually over three configurations, including the reproduced baseline, the temporally refined framework, and the standalone multimodal fusion model, and allows adjustments in performance to be traced to particular choices in the design of the system instead of the system itself. The entire assessment chain is shown in Figure. 8.

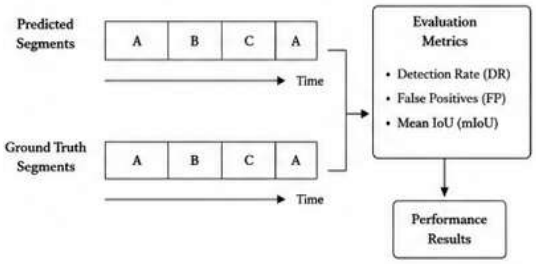

Fig 8: Evaluation Pipeline

IV. IMPROVEMENT AS PER REVIEWER COMMENTS

Every experiment has been done on Google Colab, so it can be replicated by a variety of hardware configurations and the framework can be shown to be practically executable without a powerful computing infrastructure.

This is implemented with Python 3.9 and PyTorch as the main deep learning framework, OpenCV to process video, MediaPipe to extract hand landmarks, pretrained ResNet-18 to extract visual features and neural network modules of PyTorch to implement the multimodal dual-attention architecture.

Testing is conducted on the hand gesture dataset IPN Hand Gesture Dataset, a continuous RGB gesture video dataset, with frame-level temporal annotations.

The sequential nature of its multi-gesture (its simultaneous gesture classification and its time-localization of execution) makes it especially well adapted to benchmarking online recognition systems in realistic conditions.

Three configurations are tested separately to isolate the contribution of each of the components in a transparent manner; the reproduced baseline, temporally refined framework and the standalone multimodal fusion model. All three measures of performance can be evaluated through Detection rate (DR), false positive (FP), and Mean IoU, which are all valued metrics which reflect recognition accuracy, prediction stability and temporal boundary quality.

A. Dataset Description and Experimental Set-up.
The IPN Hand Gesture Dataset is a challenging and realistic testbed to this work. In contrast to datasets of independent gesture clips, in this case, every sequence includes several gestures, executed in a natural and unstaged sequence, frame-level annotations are made to capture the type of gesture, as well as the exact start and end positions of gestures.

This renders it equally applicable in the assessment of classification accuracy and localization of a temporal boundary. Table 1 quantitatively compares the performance of the synthesiser with and without temporal refinement.

Table 1: Pre- and post-temporal performance evaluation Refinement.

Metric	Baseline System	Proposed Enhanced System
Detection Rate	0.9184	0.9288

False Positives	39	14
Mean IoU	0.7375	0.8153

OpenCV is used to extract the frames, and they are resized to 224 x 224 pixels. MediaPipe Hands recognizes 21 three-dimensional locations per frame and calculates JCD and FV representations and concatenates them into 273-dimensional skeleton feature vector.

At the same time, the 512-dimensional visual appearance vector per frame is obtained with the help of the pretrained ResNet-18. Both sequences of features are applied with overlapping sliding windows of size 16 and stride 1 to create temporally structured samples to train and make inferences.

B. Performance Evaluation Metrics.
To assess the offered framework in terms of various aspects of recognition quality, three standard measures are employed.

Detection Rate (DR) is the rate of recognizing the ground-truth gestures correctly by the system:
DR = Properly Identified Gestures / Actual Gestures.
Greater DR indicates greater gesture recognition ability.

False Positives (FP) number of predicted gesture segments which do not correspond to any real gesture occurrence. Smaller values imply higher predictive stability and a more output stream with less spurious detections.

Mean IoU is used to measure the temporal overlap of the predicted bits of gestures and their ground-truth ranges:

$$\text{IoU} = \text{Intersection} / \text{Union}$$

The increased Mean IoU values signify a more accurate localization of the temporal boundary and more reliable separation of gestures throughout the continuity recognition.

C. Detection Rate Analysis.
Even the baseline system has a Detection Rate of 0.9184, which demonstrates the usefulness of the cross-attention skeleton framework.

Using the temporal refinement module pushes this to 0.9288, with the combination of the JCD and FV features that capture hand motion and spatial relationships, ResNet-18 that adds more complementary information on appearances and the cross-attention mechanism generating more discriminative gestures.

Figure. 9 shows the comparison of the Detection Rate of all three configurations and indicates that temporal refinement is always consistent. improves detection over the baseline.

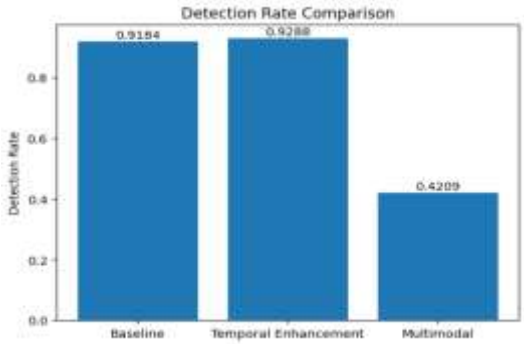

Fig. 9: Detection Rate Comparison

D. False Positive Analysis
Baseline system generates 39 False Positives, a characteristic of frame-level prediction to make noisey and fragmented predictions during continuous prediction.

This is further refined by the temporal refinement module to 14 - reduction by more than 64 (reduction) in three sequential steps: noise segment removal which removes short unstable predictions, segment merging consolidating adjacent same-class segments separated by gap, and boundary refinement which expands the boundaries of the segments to better match ground-truth intervals.

The False Positive comparison in all set ups is shown in Figure. 10, which clearly shows that the presence of temporal refinement has the greatest contribution towards the suppression of the spurious detections.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719448

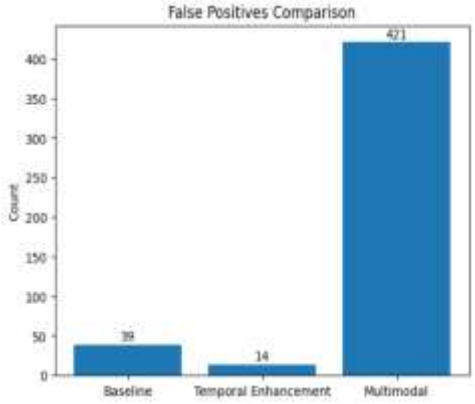

Fig. 10: False Positive Comparison

E. Mean IoU Analysis

The state-of-the-art system has a Mean IoU of 0.7375, which means that there is reasonable yet room for improvement temporal boundary alignment.

This is refined by temporal refinement to 0.8153, with most refinement coming from expanding the boundaries to cover more of the ground-truth intervals and by merging segments together to reduce temporal fragmentation. The Mean IoU comparison of all three settings is presented in Figure. 11, which validates that the quality of localization in temporal refinement has the greatest contribution.

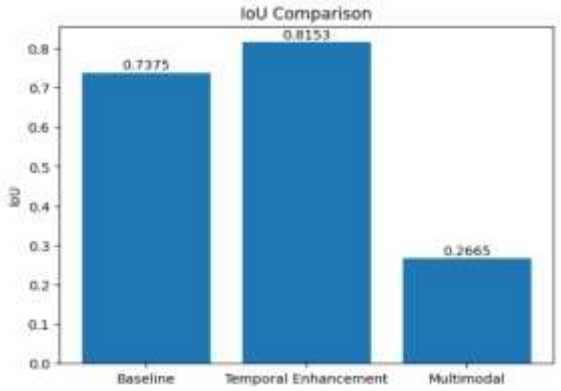

Fig. 11: Mean IoU Comparison

F. Multimodal Fusion Analysis

The introduced cross-attention fusion scheme consists of skeleton motion features, i.e., recovering to joint relationships, motion patterns, and temporal changes of said patterns with the help of JCD and FV representations, and RGB visual appearance features as the output of ResNet-18, allowing the model to use both structural and appearance-based data in the classification process. All three configurations have performance comparison summarized in Table 2.

The standalone multimodal fusion model performs poorly, compared to the temporally refined system, with the Detection Rate of 0.4209, 421 False Positives and the Mean IoU of 0.2665. This is greatly owed to three factors; the IPN Hand Gesture Dataset is mainly skeleton-based, in which motion information prevails, the additional visuals add complexity to this context that might not be entirely advantageous, and that no temporal refinement is performed after prediction resulting in the raw prediction being liable to fragmentation and noise.

However, this multimodal model shows that systemic appearance and motion fusion is possible in an architecture based on attention and it forms a strong basis to enhance future capabilities using more sophisticated fusion strategies and more visually varied training samples.

Table 2: Baseline, Temporal Refinement, and Multimodal Fusion Approaches: Performance.

Metrtic	Baseline System	Temporal Refinement System	Multimodal Fusion System
Detection Rate	0.9184	0.9288	0.4209
False Positives	39	14	421
Mean IOU	0.7375	0.8153	0.2665

G. Comparative Results Discussion

The findings in Table 1, Table 2 and Figure. As can be seen, 9-11 is an indication of temporal refinement providing the highest level of gains in all three metrics of evaluation.

Temporally refined framework is better in estimating a high value (0.9288) of Detection time, in the reduction of False Positives (39 to 14) as well as in Mean IoU (0.7375 to 0.8153), items that are indicative of a stronger continuity of gestures, cleaner predictions, and more precise localization of temporal boundary.

Although the multimodal fusion model effectively reveals the viability of the cross-modal attention-based combining system, its single-run results demonstrate the importance of the importance of post-processing (temporal) in stabilising online recognition results, which especially applies when the skeletal motion is employed as the predominant discriminative component as a skeletal motion signal.

Generally, both elements have complementary relationships in that the multimodal fusion provides the feature representation space with appearance-based context and the temporal refinement directly enhances consistency and quality of the prediction. Of the two, temporal refinement was the more productive individual refinement in the present evaluation context.

V. CONCLUSION

The paper provided a multimodal dual-attention hand gesture recognition system that combines skeletal motion features and visual appearance information through a single-deep-learning model. JCD and FV features are geometric and dynamic representations of hand movement, whereas ResNet-18 extracts complementary high-level visual representations.

Long-range dependencies in each stream are represented by independent temporal attention modules, and cross-modal attention mediates meaningful cross-modal interaction between the two streams, and the final classification layer receives the fused representation.

This framework was tested on the IPN Hand Gesture Dataset, showing that motion dynamic strength is enhanced by visual contextual information to identify gestures in more intricate continuous gestures. The temporal refinement step further refines outputs by eliminating noisy parts, merging fragmented predictions and refining gesture boundaries- yielding a higher-quality and more confident piece of recognition data on a wider range of gesture patterns, and sequence lengths.

In general, the suggested system is capable of grasping spatial and temporal features of gestures and preserves multimodal interaction with the help of focusing on attention. The findings affirm that the temporal attention coupled with cross-modal learning is a robust and effective way of doing constant hand gesture recognition with the temporal refinement component being more effective in enhancing prediction accuracy and consistency in real-world recognition settings.

VI. FUTURE SCOPE

There are a number of prospects that can be used to expand and enhance this frame. In the temporal modeling perspective, transformer-based architectures or Temporal Convolutional Networks (TCN) could be used to capture longer underlying dependencies and learn more intricate regular patterns of continuous gestures at reduced computational costs.

Another significant direction is real-time deployment. Small and slim streamlined models of inference can be used to facilitate real-world application to human-computer interaction systems, sign analysis systems and AR/VR gesture interfaces, making the framework much more usable in interactive and assistive design.

The space of multimodal input can be further enriched with additional potential, as information on depth can be added to the RGB features, and more complex visual feature extraction architectures can enhance the overall level of representation and recognition performance in extreme environments.

This temporal refinement refine stage can also be enhanced to go beyond rule-based operations and towards adaptive or learning based refine methods that directly refine the boundaries and segmentation consistency of gestures based on data, without depending on predefined thresholds.

Lastly, testing on more and more large-scale datasets with diverse individuals, gesture types, camera positions, and environments would be a more accurate assessment of the system generalizability and practical robustness and it would further confirm that the proposed multimodal attention-based framework is effective as a framework used to recognize a continuous hand gesture.

REFERENCES

[1] M. J. Chae, S. H. Han, H. Nam, J. H. Park, M. H. Cha, and S. I. Cho, "Online Hand Gesture Recognition Using Semantically Interpretable Attention Mechanism," IEEE Access, vol. 13, pp. 32329–32340, 2025, doi: 10.1109/ACCESS.2025.3540721.

[2] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai, "IPN Hand: A Video Dataset and Benchmark for Real-Time Continuous Hand Gesture Recognition," in Proc. IEEE Int. Conf. Pattern Recognit., 2021.

[3] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Online Dynamic Hand Gesture Recognition Including Efficiency Analysis," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 2, no. 2, pp. 85–97, Apr. 2020, doi: 10.1109/TBIOM.2020.2977750.

[4] P. Molchanov, X. Yang, S. Gupta, K. Kim, and J. Kautz, "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4207–4215, doi: 10.1109/CVPR.2016.456.

[5] H. Chen, X. Liu, J. Shi, and G. Zhao, "Temporal Hierarchical Dictionary Guided Decoding for Online Gesture Segmentation and Recognition," IEEE Transactions on Image Processing, vol. 29, pp. 9689–9702, 2020, doi: 10.1109/TIP.2020.3028962.

[6] S. Zhu, X. Pan, X. Cheng, and H. Guo, "Continuous Gesture Segmentation and Recognition Using 3DCNN and Convolutional LSTM," IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 1011–1021, Apr. 2019, doi: 10.1109/TMM.2018.2869278.

[7] C. Lea, M. Flynn, R. Vidal, A. Reiter, and G. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 156–165, doi: 10.1109/CVPR.2017.113.

[8] X. Yang, P. Molchanov, and J. Kautz, "Making Convolutional Networks Recurrent for Visual Sequence Learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6469–6478, doi: 10.1109/CVPR.2018.00677.

[9] H. Wang, P. Wang, Z. Song, and W. Li, "Large-Scale Multimodal Gesture Recognition Using Heterogeneous Networks," IEEE Transactions on Cybernetics, vol. 49, no. 10, pp. 3660–3673, Oct. 2019, doi: 10.1109/TCYB.2018.2844809.

[10] M. Abavisani, H. Joze, and V. M. Patel, "Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition with Multimodal Training," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1165–1174, doi: 10.1109/CVPR42600.2020.00124.

[11] R. Kopuklu, N. Kose, A. Gunduz, and G. Rigoll, "Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks," in Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., 2019, doi: 10.1109/FG.2019.8756576.

[12] L. Qu, H. Wu, T. Yang, L. Zhang, and Y. Sun, "Dynamic Hand Gesture Classification Based on Multichannel Radar Using Multistream Fusion 1-D Convolutional Neural Network," IEEE Sensors Journal, vol. 22, no. 24, pp. 24083–24093, Dec. 2022, doi: 10.1109/JSEN.2022.3216604.

[13] W. Zhang, J. Wang, and F. Lan, "Dynamic hand gesture recognition based on short-term sampling neural networks," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 1, pp. 110–120, Jan. 2021, doi: 10.1109/JAS.2020.1003465.

[14] D. Zhao, H. Li, and S. Yan, "Spatial-Temporal Synchronous Transformer for Skeleton-Based Hand Gesture Recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 3, pp. 1403–1412, Mar. 2024, doi: 10.1109/TCSVT.2023.3295084.

[15] Y. Cao, J. Li, C. Chakraborty, L. Qin, L. Tao, and X. Shao, "Temporal Segment Neural Networks-Enabled Dynamic Hand-Gesture Recognition for Industrial Cyber-Physical Authentication Systems," IEEE Systems Journal, vol. 17, no. 4, pp. 5315–5326, Dec. 2023, doi: 10.1109/JSYST.2023.3306380.

[16] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, doi: 10.1109/CVPR.2019.00132.

[17] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, doi: 10.1109/CVPR.2019.00054.

[18] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition with Directed Graph Neural Networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, doi: 10.1109/CVPR.2019.00810.

[19] Y. Du, W. Wang, and L. Wang, "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1110–1118, doi: 10.1109/CVPR.2015.7298714.

[20] H. Wang and L. Wang, "Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 499–508, doi: 10.1109/CVPR.2017.61.

[21] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 143–152, doi: 10.1109/CVPR42600.2020.00022.

[22] X. Chen, Y. Ye, and J. Xu, "Richly Activated Graph Convolutional Network for Robust Skeleton-Based Action Recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 1915–1925, May 2021, doi: 10.1109/TCSVT.2020.3015051.

[23] Z. Tu, H. Zhang, H. Liu, J. Yuan, and J. Li, "Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 1474–1488, Feb. 2023, doi: 10.1109/TPAMI.2022.3157033.

[24] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan, and A. C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, pp. 2684–2701, Oct. 2020, doi: 10.1109/TPAMI.2019.2916873.

[25] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A Large-Scale Dataset for 3D Human Activity Analysis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1010–1019, doi: 10.1109/CVPR.2016.115.

[26] J. Y. Kim, B. H. Kang, and D. H. Im, "Gate-Shift Networks for Video Action Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, doi: 10.1109/CVPR42600.2020.01245.

[27] D. Tran et al., "Learning Spatiotemporal Features With 3D Convolutional Networks," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.

[28] T. Zhang, W. Zheng, Z. Cui, C. Shan, and J. Yang, "A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training," IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1880–1891, Jul. 2019, doi: 10.1109/TMM.2018.2889563.

[29] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 9, pp. 1685–1699, 2009, doi: 10.1109/TPAMI.2008.203.

[30] A. Osman Hashi, S. Zaiton Mohd Hashim, and A. Bte Asamah, "A Systematic Review of Hand Gesture Recognition: An Update from 2018 to 2024," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3421992.

[31] Z. R. Saeed, Z. B. Zainol, B. B. Zaidan, and A. H. Alamoodi, "A Systematic Review on Systems-Based Sensory Gloves for Sign Language Pattern Recognition: An Update from 2017 to 2022," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3219430.