Vol. 08, No. 2 (2026) 1389-1398, doi: 10.24874/PES08.02B.027

Proceedings on Engineering Sciences

www.pesjournal.net

TWO-STAGE FRAMEWORK FOR CONTINUOUS HAND GESTURE SPOTTING AND RECOGNITION USING 3D SKELETON JOINT KINEMATICS

D. V. Srihari Babu
Y Narasimha Reddy
M Nagaraju
Bakkala Santha Kumar[1]

Received 27.11.2025.
Revised 18.03.2026.
Accepted 02.04.2026.

Keywords:

Continuous hand gesture recognition,
Gesture spotting,
Temporal localization,
Skeleton kinematic descriptor,
Skeleton-based action recognition

A B S T R A C T

Continuous hand gesture recognition needs not only the spotting (temporal localization) but also the gesture identification (classification) under the ambiguity of transitions and motion variation. This paper introduces a two-stage, cloud-deployable framework that somehow performs gesture spotting based on a compact 3D skeletal kinematic descriptor, and on the other hand, classifies isolated gestures based on key joint trajectories. We perform a comparative ablation study for three spotting variants, namely BiLSTM+BCE, BiLSTM+weighted BCE and attention+focal loss, and report the frame accuracy and mean Jaccard overlap, followed by a skeletal LSTM classifier accuracy. In controlled continuous stream experiments, the best spotting variant has 85.07% frame accuracy, 67.35% of mean Jaccard, and 94.43% of classifier accuracy. Finally, we address the scalability and deployment by the proposed model export for the cloud inference and the evaluation on ChaLearn LAP ConGD and the representation pretraining using Human3.6M as future works.

© 2026 Published by Faculty of Engineering

1. INTRODUCTION

Hand gesture recognition has become an integral part of natural human-computer interaction and has applications that range from virtual and augmented reality, robotics, intelligent surveillance, healthcare, and smart environments. In many practical scenarios, gestures are done in continuous manner, in which meaningful gestures are interleaved with non-gesture transition movements. This is a challenging setting against isolated gesture recognition, as systems are required to not only recognize categories of gestures but also to locate gesture boundaries precisely at their time of use. Continuous gesture recognition is thus often split into two closely coupled problems: gesture spotting, which attempts to identify the temporal beginning and ending points of the gestures in an un-segmented stream of a motion, and gesture classification, which attaches semantic labels to the detected segments. Among these, the more difficult task is still spotting, since transition frames are characterized by often very subtle motions that are very similar to gesture frames. Errors in spotting may carry over to the classification stage making robust temporal localization a crucial requirement for dependable continuous recognition systems.

[1] Corresponding author: Bakkala Santha Kumar
Email: santhakumar.bakkala@gmail.com

Large-scale benchmarks including ChaLearn LAP Continuous Gesture Dataset (ConGD) have made a great leap forward on the research of this area, providing realistic RGB-D recordings with temporal gesture boundaries. However, these methods based on high-dimensional RGB or depth features sometimes have severe computer cost and privacy issues, restricting their incrustation to in a real time and cloud-based sera. In contrast, the skeleton-based representations provide a compact and privacy-preserving method, where important motion dynamics are captured, with a lower data dimension. Recent improvements in skeleton-based action and gesture recognition have shown that recurrent neural networks, graph neural networks and transformers architectures are effective for modelling spatiotemporal joint dynamics. Nevertheless, most of the existing works are mostly concerned with isolated recognition or approach continuous recognition in an end-to-end fashion without explicitly analyzing the spotting and classification stage's contribution. In addition, not much work has been put into the problem of lightweight spotting representations that are efficient to deploy in the cloud or edge and do not sacrifice much temporal localization performance.

From a practical point of view, an effective continuous gesture recognition system should meet four important requirements:

- Precise temporal localization of the boundaries of gestures in transition ambiguity,
- High accuracy of recognition of isolated gestures,
- Computational efficiency and scalability for cloud-based deployment, and
- Reproducibility and interpretability allow systematic evaluation and ablation to be conducted.

Motivated by these requirements, this work examines whether a skeletal kinematic descriptor that is now prominent and well-defined and is burrowed down to a minimal number of joints throughout the hand, can afford good gesture spot and recognition, if combined with suitable temporal modeling and post-processing strategies. This paper presents a cloud-ready two-stage framework on continuous hand gesture recognition using 3D skeletal joint kinematics only. The novelty here is to develop gesture spotting based on a unified hand-finger speed descriptor with additional temporal change measured as $[v_k, \Delta v_k]$, this includes both global hand motion and fine-grained finger articulation using only 3 key joints. In contrast to a high dimensional representation of visual features, this representation is compact, efficient and appropriate for scalable inference. Furthermore, the work introduces systematic comparative work of different spotting exposures of standard binary cross-entropy, weighted binary cross entropy, attention enhanced focal loss in identical feature representations and post processing conditions. An optional boundary refinement module is also introduced to produce stable temporal localization at negligible computational overhead. Together, these aspects give both methodological knowledge and expert designing knowledge for continuous skeleton-based gestures recognition systems.

The major contributions of this work are listed as follows:

- A set of lightweight skeleton kinematic miscellaneous associations for unifying glob hand speed and finger articulation dynamics combined significant and with such temporal figures for change to robust gesture spotting purposes.
- A modular two-stage architecture that separates the gesture spotting and classification operations allowing the temporal localization and recognition performance to be optimized and analyzed independently.
- A large-scale comparative ablation study of 3 spotting variants in terms of frame-wise accuracy and mean Jaccard overlap with and without refining the boundaries.
- Cloud deployable implementation that is exported as TensorFlow SavedModels, having support for scalable inference and integration with state-of-the-art serving infrastructures.

The rest of the paper is structured as follows. Section II is a review of relevant literature on continuous gesture recognition and skeleton models. The formulated materials and methods are discussed in section III, where they comprise feature extraction, network architecture, and training strategies. The IV has experimental results and discussion with comparative analysis of spotting variants and classification performance. Finally, Section V concludes the paper and presents ideas for future work.

2. LITERATURE REVIEW

Continuous gesture recognition involves detecting the appearance of sparse gesture events from long and untimed streams while at the same time not detecting false recognition from transition motions. Early deep learning solutions had explicitly taken the spotting problem as an online temporal localization problem. Benitez-Garcia et al. proposed the multi-stream recurrent architectures for detecting finger gestures in long sequences, proving that temporal modeling and cues of hand motion are key in suppressing the natural movements that are similar to target gestures (Benitez-Garcia et al., 2020). In a related work in the proceeding of the Experiments on Audio, Speech and Music Processing, the same line of research focused on segment-level modeling by exploiting similarity between the start frame and the end frame to enhance the continuous spotting robustness in realistic interactive environments (Benitez-Garcia et al., 2022). The development of data sets and testing procedures has had an influence as well. LD-ConGR was proposed as a large

RGB-D benchmark with particular emphasis on long-distance continuous gesture recognition, where gesture regions are small compared to the background and segment durations are very diverse - all of which are very similar to the scenarios seen for smart-room implementations (Liu et al., 2022). These research papers are all pointing to the fact that the issue in spotting performance lies in (i) the ability to perform very strong temporal modelling at changing duration, and (ii) Boundary-stabilization techniques that minimize fragmentation and jitter in respective saved passages. A recent survey by Emporio et al. showed the systematization of the benchmarks and measures to assess continuous hand gesture recognition, underlining the importance of a fair evaluation that has to consider the simultaneous reporting of frame-wise measures and and segment overlap measures (e.g., IoU/Jaccard), as it can happen that the evaluation of the temporal accuracy and the temporal accuracy is really different from the evaluation of the classification accuracy in case of long streams (Emporio et al., 2025). This has the direct motivation to report both spotting metric (frame accuracy, Jaccard /Iou) and recognition metric in a two-stage pipeline.

The skeleton-based recognition has attracted momentum because of its compactness, insensibility to change in illumination and appearances and representations that are privacy-friendly. The spatial-temporal graph modeling (ST-GCN-style designs) provided a sound starting point in that joints are modeled as nodes in the graph and learning of a temporal and spatial reliance are jointly executed. Later studies highlighted that ongoing enhancement of performance is usually due to improvement in topology learning, improvement in temporal modules rather than simple increment in depth. SBI-DHGR also presented a dynamic hand gesture recognition system based on skeletons and explicit palm-centered modeling and augmentation policies in hand-gesture, where results with strong recognition are obtained and with low computational costs by being careful in joint choice and gesture-driven preprocessing (Narayan et al., 2023). This temporary can be used with lightweight joint-based pipelines in cases when real-time inference and execution on the cloud are desired. One major shortcoming of skeleton GCN early skeletal models is the application of fixed or weakly adaptation structures of adjacency. CTR-GCN proposed refinement of channel topology, meaning the model could learn the various topological affinities each channel with its features contributing to the skeletons and this approach learned a significant performance with a small overhead (Chen et al., 2021). This approach is one of the largest trends: adaptive topology learning to learn action-specific joint co-occurrence patterns. Complementary work considered feedback to infuse upper levels of semantics into lower levels. Feedback Graph Convolutional Network (FGCN) added feedback connections and thus overcomes the feedforward barriers because the main task is to enhance progressive extraction of spatiotemporal features (Yang et al., 2022). Recently, feedback-directed and pose-correction versions are also following this trend, which implies that topology learning, correction and feedback fusion technology can become strongly robust to pose noise and domain shifts (Ran & Yang, 2025; Gao et al., 2025).

Transformers made skeleton modeling worldwide in the attention of oneself to capture long-range joint interactions and time constraints. ST-TR (Spatial-Temporal Transformer Networks) had inter-frame (temporal) and intra-frame (spatial) attention operators to model relationships in models, which produced high results on mainstream skeleton benchmarks and informed the design of subsequent transformer-based skeleton models (Plizzari et al., 2021). In the low-latency setting Action Transformer (AcT) demonstrated that fully self-attentional models could also scale better than high greater recurrent/convolutional designs when have been applied on short time-pitfalls, anchoring a valuable design choice in edge-cloud split inference (Mazzia et al., 2022). An exclusive review offered by Xin et al. gave an overview of the transformer methods of skeleton action recognition based on key patterns, including multi-stream inputs (joints/bones/velocity), efficient attention, and a combination with graph structures to sacrifice accuracy with computing (Xin et al., 2023). Parallel Hypergraph Transformer (Hyperformer) and design Hypergraph-attention the fixed graph topology was ultimately overcome by enabling richer joint grouping and higher-order relationships with competitive performance and interest in attention-graph hybrids (Zhou et al., 2022; Ray et al., 2025). Although the full-body action is aimed at by the large number of skeleton models, gesture recognition presents some exclusive requirements: subtle movements, brief motion, and transitions. Temporal Decoupling GCN Temporal Decoupling GCN (TD-GCN) explicitly overcame the drawback of using a single adjacency matrix over time, instead using different adjacency structure at each frame and thus offering more flexibility to model Gesture-like temporal dynamics and better modeling of temporal variation (Liu et al., 2024). This is especially applicable to continuous spotting pipelines whereby the patterns of boundary frames in most cases have different topologies as compared to peak gesture frames.

Continuous abnormal activity and gesture datasets require a lot of time and money to be labeled. Skeleton representation learning under supervision of self has thus been increasing speedily. CRRL introduced learning in contrast-reconstruction so that the posture and motion could be encoded together in an unlabeled way, and the practicality of learning in useful motion representation without supervision was revealed (Wang et al., 2022). Other self-supervised directions, such as cross-stream contrastive learning, point out enhanced uses of multi-view or multi-stream signals beyond signal contractiveness to single streams augmentation (Li et al., 2023). Motivated by such results, pretraining (self-

supervised or large scale supervised) is used to improve downstream spotting/classification tasks. Survey articles are becoming more superimposing of trends in methods and practice in evaluation, as the discipline reaches maturity. Ren et al. 2024 offered a global overview of RNN, CNN, GCN, and transformers in 3D estimating skeletal-based action recognition and presented datasets and standard experimental guidelines (Ren et al., 2024). In the case of continuous gestures, Emporio et al. focused on the benchmark taxonomies and the relevance of stable spotting measures. These questionnaires are one of the reasons why the methodological rigor agreed upon in this work has been pursued: separation of spotting and classification and reporting of both the temporal overlap and recognition measures (Chen et al., 2025; Chi et al., 2025; Zhu et al., 2023; Cui et al., 2025).

Although considerable advances have been made in skeleton-based recognition based on graph and transformer architectures, most recent algorithms are tested on truncated sequences and focus on classification accuracy and not the ability to localize temporally in continuous streams. The ongoing research of gesture recognition illustrates that detection of quality may widely differ compared to classification, and to implement a uniform kind of reporting, metrics of frame and segment overlap are required. Besides, RGB-D pipelines can be run at high-dimensional numbers, which limits the computational and privacy capabilities, hence making them difficult to scale. Inspired by these voids, the current work uses a two-step spotting-classification framework and proposes a small-sized kinematic representative obtained by light 3D joint measurements, which allow finding boundaries and cloud-cost-efficient inference and provides capability to have systematic ablation of various spotting losses and refinement of boundary methods.

3. RESEARCH METHODOLOGY

This paper suggests a two-stage system of hand gesture recognition in permanent mode by utilizing information about 3D skeletal joints. The model breaks down the problem into gesture spotting that localizes meaningful segments of gesture motions in continuous streams of motion, and gesture classification that is used to provide semantic labels on the isolated segments. This decomposition allows the independent optimization and assessment of localization of time and recognition accuracy which is very important in continuous gesture environments in which the transition moves tend to blur gesture boundaries. Depending on its deployment in clouds, the entire pipeline will be light and modular. Figure 1. Generalization of the suggested two-stage continuous gesture recognition system. The incoming skeletal motion streams are initially handled through a small- scale kinematic gesture spotting describer, segmentation, and refinement, and eventually through gesture recognition network based on an LSTM.

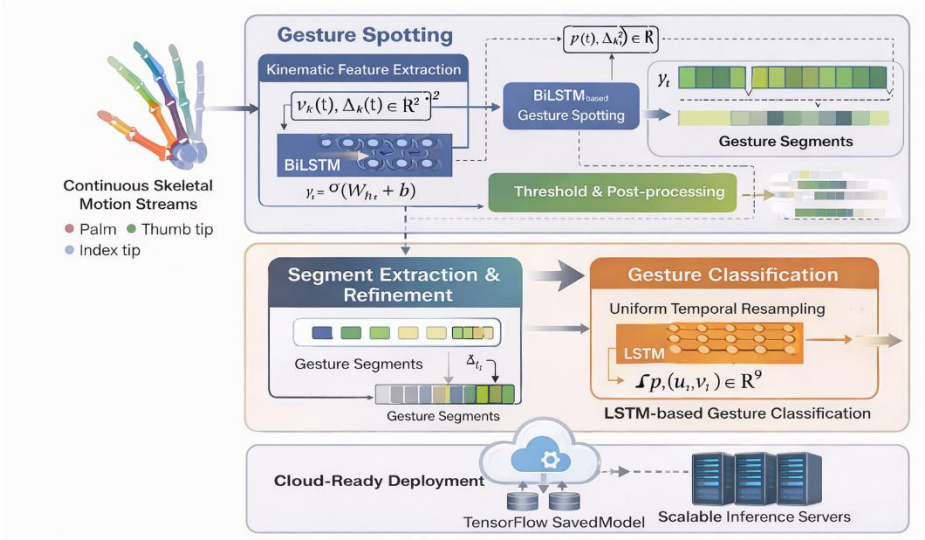


Figure. 1 Overall framework architecture

3.1 Data Availability

The suggested framework is aimed at working on large-scale continuous gesture data, including ChaLearn LAP Continuous Gesture Dataset (ConGD), where gestures are annotated with temporal frames and denoted with gesture labels. ConGD is a universal continuous gesture recognition benchmark, and it allows to assess both spotting and classification performance in real-life scenarios. The use of ConGD is, however, licensed and registered as required by certain regulations. Moreover, the framework facilitates representation learning on large scale 3D pose datasets like Human3.6M which offers high-precision 3D joint coordinates which are taken through multi-view motion capture systems. The pretraining and generalization of skeletal representations is often done using Human3.6M. Because of the limitations with respect to access of ConGD during the experiment, a synthetic continuous dataset under controlled conditions was synthesized to test the

proposed approach and to allow systematic ablation analysis. The synthetic data recreates real continuous gesture streams by interconnecting several segments of gestures with transition periods and generating related 3D joint paths. Its benefit is complete reproducibility and control of gesture duration, transition duration and variability of motion that prove very helpful in the assessment of temporal segmentation quality.

The motion stream is encoded with the 3D position of three prominent hand joints as the palm, thumb tip and index fingertip. The joints have been identified to provide a balance between representational compressiveness and discriminative power. Whereas the palm joint records hand motion worldwide, the relative positioning of the thumb and index finger records fine-grained finger articulation, which is required to differentiate gestures of similar overall movement patterns. The joint coordinates at times *t* be denoted as:

$$P_t \in \mathbb{R}^3\ (palm), u_t \in \mathbb{R}^3\ (thumb\ tip),\ v_t \in \mathbb{R}^3\ (index\ tip)$$

3.2 Kinematic Feature Extraction

To allow powerful gesture recognition based on continuous skeletal examples streams, kinematic properties are obtained by a small number of joints of the hand that represent both large-scale motion and minute finger motions. Overall hand motion is encoded in the palm joint, finger motion at the fingers level is represented by the thumb-tip and index-tip joints, which are important in differentiating gestures with a similar gait of overall motion. To measure the global hand movement in every frame t, the Euclidean distance between the palm joints of the two frames is computed and multiplied by the sampling frequency to get a measurement of velocity. Simultaneously with these, Finer finger articulation is modelled by the temporal fluctuation of the Euclidean distance between the thumb tip and the index fingertip, which contains opening-and-closing movements which often describe hand movements. The two motion components are then synthesized into a single kinematic measure, denoted $\boldsymbol{v_k(t)}$ that represents both translational hand motion and relative finger motion within one scalar signal. To increase some more sensitivity to sudden changes in motion that are often seen in the body at the edges of gesture initiation and exit, a temporal difference signal Δvk(t) is calculated as the first derivative of the combined speed signal. The feature vector that results $[\boldsymbol{v_k(t)}, \boldsymbol{\Delta v_k(t)}]$ is a succinct, yet highly expressive, description of hand motions, allowing an effective temporal model, and well-adapted to real-time and cloud-based implementation.

3.3 Gesture Spotting Network

Gesture spotting is a frame-based binary classification problem in which every frame in a continuous sequence of motion is perceived as probability of either being a part of a meaningful gesture, or transition period. Given as an input an input sequence of kinematic feature vectors $X = \{x_1, x_2, \ldots, x_L\}$, the spotting module aims at estimating the posterior probability of each frame $P(y_t=1 / X)$. The Bidirectional Long Short-Term Memory (BiLSTM) network is used to provide long-range temporal dependencies and context on the frame boundaries of a gesture. A Bidirectional Long Short-Term Memory (BiLSTM) network to capture long-range dependencies in gesture boundaries and the contextual information. The BiLSTM has multiple benefits in this regard because processing the sequence both forward and backward interoperates past and future cues of motion that is particularly critical in a continuous gesture environment where boundary frames tend to rely on both previous and subsequent motion patterns. The BiLSTM outputs are run through fully connected layer with the sigmoid activation to generate frame-wise probability of gestures. To research how varying learning strategies affect the process of localizing the boundaries, three variants of spotting are compared: a control variant of BiLSTM trained on the standard binary cross-entropy loss, a variant with class-imbalance sensitivity on weighted binary cross-entropy, and an improved variant with a self-attention mechanism trained on the focal loss which highlights hard-to-classify boundary frames. With this design, a systematic comparison between the loss functions and architectural complexity are possible when features are represented using the same set of features.

3.4 Temporal Smoothing and Segment Extraction

The framewise likelihoods generated by the spotting network could have short term variation because of noise or unclear motion patterns at gesture boundaries. To enhance the temporal stability, a sliding-window moving average operation on the predicted probabilities is applied to achieve temporal stability. This is a high-frequency noise-reduction halting, which leaves the structure of gesture intervals intact. Gesture segments are then born by using a fixed probability value on the smoothing signal which will leave behind continuous blocks of frames that correspond to gesture activity. To make it even stronger, small pauses between neighboring gesture segments are eliminated, and those shorter than the specified minimum duration are removed. The following post-processing steps are necessary to make sure that the extracted segments are related to meaningful gestures instead of unrelated activations, which are input cleaner to the next classification step.

3.5 Boundary Refinement Strategy

Even though meaningful gesture intervals are obtained with threshold-based segmentation, inaccuracies in boundary determination can still occur because of a smooth transition of motions or ambiguous estimates on a frame-by-frame basis. To solve this problem, a lightweight post processing step of optional boundary refinement strategy is proposed. The refinement process

makes use of hysteresis-like expansion principles, by which the identified gesture portions can be expanded to adjoining frames with a high confidence rating. The segments that are adjacent or overlapping are then joined together to remove the fragmentation. This refinements technique enhances pre-estimation with ground-truth gesture boundaries regarding time utility without raising model complexity on information inferences and reaction time. This makes boundary refinement especially appropriate in the case of real time and cloud-based deployment, where computing efficiency is paramount.

3.6 Gesture Classification Network

After gesture parts are localized, each part is subjected to gesture classification network. Since the detected segments have different lengths, all segments are resampled to a common time scale length to provide a uniform input representation. In the case of each resampled frame, the 3D coordinates of the palm, thumb-tip and index-tip joints are concatenated to create a small spatial description. The resulting sequence is then inputted into a Long Short-Term Memory (LSTM) network itself, that predicts the time basis dependencies within the isolated gesture itself. The last latent feature undergoes a fully connected layer using softmax activation to generate probability of gesture classes. With the release of classification and spotting, the classifier can concentrate on discriminative gesture dynamics and classifies high recognition rates after consistent localization of time is attained.

3.7 Training Strategy and Evaluation Metrics

The entire parts of the suggested architecture were applied with the help of the TensorFlow and Keras deep learning packages. The spotting and classification networks were trained separately to enable them to be judged fairly on each step. They used series of early stopping and adaptive learning-rate scheduling in training which helps in overfitting and stable convergence. The loss-function design was used directly to overcome class imbalance in spotting task instead of being ruthless on data resampling. At the end of the training process, models were stored in their native Keras format to aid reproducibility and created models to aid cloud-inference, load balancing, scalable deployment with current serving infrastructures. The complementary metrics that measure the quality of localization and recognition of time were used to evaluate the performance of the proposed framework jointly. The evaluation of gesture spotting performance was performed based on frame-wise accuracy and mean Jaccard overlap, both of which evaluate point-wise accuracy and temporal resolution of segments between segment predicted and ground-truth gesture ranges. The overall performance in terms of gesture classification was measured on both the basis of overall accuracy and macro-averaged precision, recall, and the F1-score to take into consideration the variation in performance per class. Synergistically, these measures have a full description of the framework effectiveness in continuous gesture recognition situations.

4. EXPERIMENTS AND RESULTS

Every experiment was carried out with the help of the TensorFlow/Keras platform on a continuous skeletal movement dataset. The experimental design is based on the evaluation protocol of two stages, (i) assessment of gesture spotting in continuous stream, and (ii) isolated assessment of gesture classification (using isolated segments). This division can enable the accurate characterization of the quality of temporal localization and recognition that is important in continuous gesture conditions. The data is a sequence of long motion patterns, and these patterns are made up of several segments of gestures mixed with transition periods. All the frames are described in the proposed kinematic feature based on 3D hand joints. Training and test subsets are separated by 80:20 ratio in every sequence avoiding temporal leakage. Early stopping and learning-rate scheduling were used to train all models so that their convergence would be stable. There were three spotting variants that were executed and tested under the same conditions:

- E1: BiLSTM was trained using standard binary cross-entropy (BCE) only.
- E2: BiLSTM and weighted BCE to overcome the imbalance in classes,
- E3: Augmented model BiLSTM with self-attention and trained on focal loss.

To classify gestures, a single LSTM-based classifier was trained using uniformly resampled gestures segments obtained by using the ground-truth scales. This classifier was held constant in all spotting variants in case of fair comparison.

Future research will work towards working out extension of the proposed framework with real world benchmark data sets, particularly the ChaLearn LAP Continuous Gesture Dataset, following official evaluation protocols. In addition, we will explore representation pretraining on large-scale motion capture datasets such as Human3.6M to improve generalization on subjects and styles of gestures. Further extensions consider multimodal fusion with RGB or depth cues, adaptive thresholding strategies to spot, and considerations as to deploy spotting in edge-cloud collaborative environments. Two complementary measures were used to measure gestures spotting performance. Frame-wise accuracy: This is the ratio of correctly identified frames in the sequence as a whole, whereas mean Jaccard overlap: this is the consistency between the frame interval that predicts the prediction and the frame interval that contains the ground-truth gesture sites, assessing the quality of localization of the boundaries. To determine the performance of classes using gesture classification, the overall accuracy was evaluated and was reinforced with the analysis of

confusion matrices to recognize the behavior of classes. Good macro- and weighted-average values have shown finding of uniformity in performance of classification among classes.

Table 1. Gesture Classification Results

	Precision	Recall	F1-Score	Support
0	0.8671	0.7965	0.8303	172
1	1.0000	1.0000	1.0000	162
2	1.0000	1.0000	1.0000	166
3	1.0000	1.0000	1.0000	169
4	0.8045	0.8727	0.8372	165
5	1.0000	1.0000	1.0000	171
Accuracy			0.9443	1005
Macro Avg	0.9453	0.9449	0.9446	1005
Weighted Avg	0.9452	0.9443	0.9442	1005

Table 1 provides the precision of the proposed LSTM-based gesture classifier per class including recall and F1-score when tested on individual skeletal gesture segments. The overall accuracy of classification by the model is 94.43, which means that a significant percentage of gestures in response to the entire test is correctly identified. The macro-averaged precision (0.9453) and recall (0.9449) are good indicators of the fact that the classifier does not show an uneven performance across the gesture categories, and the fact that the macro-averaged and weighted-average metrics are close enough to each other indicates that the distribution of classes does not create an appreciable bias in the analysis. Class-by-class analysis indicates that gesture classes 1, 2, 3, and 5 achieve high precision, recall and F1-score (1.0000) indicating an absolute differentiation between these types of gestures under the skeletal representation adopted. Comparatively, gesture classes 0 and 4, with F1-scores of 0.8303 and 0.8372, respectively, are motivated by lower levels of precision or recall.

Such deviations suggest that there are inherent local classification ambiguity and not systematic failure which probably has to do with overlapping kinematic properties of these gestures. Notably, the F1-scores of all classes remain larger than 0.83, which proves that classification errors are not that large and they are concentrated in a small part of classes. Statistically, the minimal spread in the macro-averaged and weighted-average metrics, as well as equally high performance in all the classes, suggests that the classifier can generalize a fair deal in the gesture category. Such findings prove that with proper temporal segmentation, the skeletal representation and LSTM classifier offer a sound and statistically consistent system to be used in gesture recognition.
Table 1 demonstrates the performance of the various classes of gestures in a class-wise manner that is further reinforced by the confusion matrix presented in Figure 2, which demonstrates that the misclassification attempt is by a limited number of similar gesture classes. Table 2 sums up the frame-wise accuracy of spotting the three variants. The frame accuracy of the best baseline E1 (BiLSTM + BCE) was 85.07 which shows that their overall ability to discriminate between gesture and transition frames was high. The weighted BCE model (E2) was also able to give an accuracy of 80.97, whereas the attention enhanced focal loss model was able to get an accuracy of 82.33. The above results indicate that using standard BCE gives a superior prediction of frame-level probabilities compared to imbalance-aware loss and hard-sample loss, which can focus on the ambiguous transition frames too much.

Table 2. Gesture Spotting Performance Comparison

Model Variant	**Frame Accuracy**	**Mean Jaccard**	**Mean Jaccard (Refined)**
E1: BiLSTM + BCE	**85.07%**	**0.6735**	0.6716
E2: BiLSTM + Weighted BCE	80.97%	0.6330	0.6203
E3: Attention + Focal Loss	82.33%	0.5908	0.6030

Segment-level localization accuracy was determined by use of means of Jaccard overlap. E1 again made the best score with a Jaccard score of 0.6735, and thus, better temporal overlap with gesture intervals of ground-truth. The Jaccard score of E2 was 0.6330 and E3 was scored at 0.5908 with a bit more boundary fragmentation but with a moderate level of frame accuracy. In further stabilizing prediction of the boundaries, a refinement strategy of the boundaries was used as an option. This led to E3 and E1 and E2 having slight changes but E3 had a significant change in Jaccard overlap (0.6030) after refinement. This finding shows that attention-based spotting models are more likely to generate sharper yet fewer stable boundaries that are useful in post-processing. Simultaneously, even the baseline BiLSTM already forms quite smooth and uninterrupted portions. In general, the trained baseline BiLSTM using the standard BCE shows the most reasonable trade-off between simplicity and temporal localization accuracy.

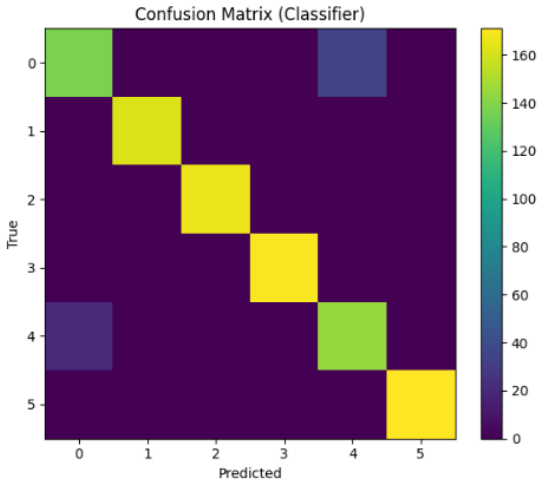
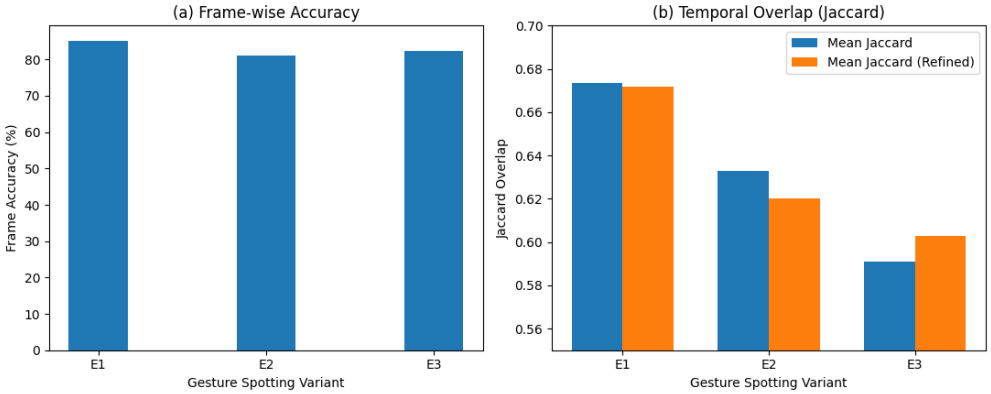


Figure 2. Confusion Matrix of Gesture Classifier

Both the complexity of architecture and changes to the loss function do not always correspond to better spotting behavior especially in continuous motion streams with slow transitions. Isolated segmentation was done with palm, thumb-tip, and index-tip joint coordinate sequences through an LSTM network that was trained on resampled sequences. The overall accuracy of the classifier was 94.43 which is a strong discriminative performance when gesture limits are known. Based on the confusion matrix, most of the classes of gestures are identified with almost perfect accuracy. Most of the misclassifications are concentrated between two synthetically defined classes sharing similar patterns of hand motions in the globe, which showed that the ambiguity is due mainly to the fine-grained articulation of the fingers, and that there are not as many time constraints due to time modeling aspects. The outcomes prove that the suggested skeletal representation is very useful during gesture recognition in case the segmentation is correct. Figure 2. Confusion of the LSTM-based gesture classifier on isolated skeletal gesture segment. A high degree of diagonal dominance suggests a high degree of recognition with little confusion between similar classifications of gestures that can be kinetically similar.



(a) (b)

Figure. 3. Performance of comparative gesture spotting with three model variants: (a) Frame-wise performance and (b) Mean Jaccard overlap pre- and post-refinement on boundary

Figure 3. depicts the comparison of three variants of models based on the concept of frame-wise accuracy, mean Jaccard overlap, and refined mean Jaccard overlap. The baseline BiLSTM which uses binary cross-entropy (E1) has the best temporal localization performance, whereas variants which use weighted and focal-loss have less overlap consistency.

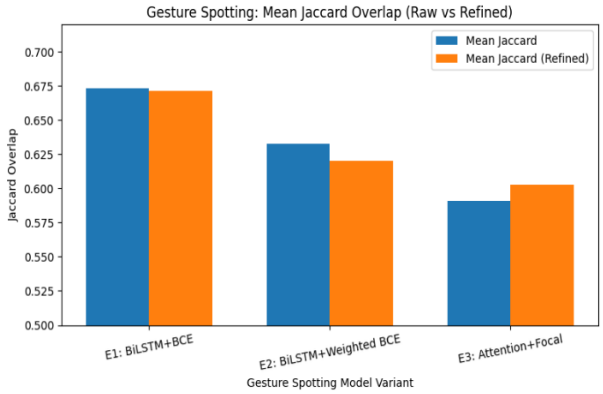


Figure. 4 Mean raw vs. refined gesture spotting Jaccard overlaps between variants of the model

Figure 4 shows the average Jaccard overlap obtained with all three variants of the gesture spotting along with the variants after boundaries are refined. The highest Jaccard overlap is achieved with E1 (BiLSTM + BCE) both raw and refined, meaning that it better predicts the temporal coverage of the gesture segments between predictions and ground-truth in both cases. The refined Jaccard score of E1 is also almost the same and it indicates that this model naturally generates smooth and constant gesture limits with least fragmentation. E2 (BiLSTM + Weighted BCE) variant demonstrates a significant decrease in the overlapping of the Jaccard between E1 and a boundary refinement slightly worsens its performance. This behavior indicates that reweighted loss function to correct class imbalance can also add some extra uncertainty to boundaries, which leads to a less competitive segment localization. Contrarily, E3 (Attention + Focal Loss) shows the raw Jaccard overlap and has an observable improvement upon the refinement of its boundaries. This shows that although the attention-based model with focal loss results in sharper yet fragmented boundary predictions, primarily post-processing can merge gesture segments, as well as enhance temporal overlap. All in all, baseline BiLSTM models yield more reliable localization, but attention-based models can better utilize explicit boundary refinement.

5. DISCUSSIONS

Several important points stand out in the results of the experiment. First, gesture spotting is still the most defining bottleneck in the continuous gesture recognition pipelines since the classification accuracy is far higher

than the time-localization accuracy. Second, although loss functions based on class imbalance or hard sample mining are particularly theoretically attractive, they can be disadvantageous to consistency in practice when transition frames overwhelm the time stream. Third, some lightweight post-processing methods (like refining the boundaries) can reduce the problem of the boundary instability to some extent, especially in attention-based models. Regarding deployment, the findings indicate that a small kinematic representation, together with BiLSTM spotter, will be able to maintain high localization rates without using high-dimensional visual attributes. The model that has been exported also allows inference with load-balancing and cloud-based serving infrastructures that scale.

6. CONCLUSION

This paper described a two-stage, cloud-ready architecture for continuous spotting and recognition of hand gestures by using compact 3D skeletal joint kinematics. By splitting the problem into gesture spotting and gesture classification, the proposed approach allows specifically modeling the temporal localization and semantic recognition, which is important for continuous gesture scenarios where transition ambiguity is a major problem. A light-weight kinematic descriptor that combines the global hand motion and finger articulation with the augmentation of temporal change was proposed for the spotting of the gesture. Experimental results show that the frame-wise accuracy and Jaccard overlap of the temporal localization of the spotter trained with standard binary cross-entropy, respectively, are the most stable 85.07% and 67.35%. In contrast, weighted and focal versions of the loss, while very attractive from the perspective of how to address imbalance, were found to have fewer stable boundaries under the conditions tested. For gesture classification, an LSTM working on fixed-length sequences of the palm joint and index finger thumb tip joint and index fingertip joint showed 94.43% recognition accuracy, displaying the discriminative force of the compact skeletal models once the boundaries of gesture are identified. The comparative analysis has an important insight - namely that gesture spotting is the main bottleneck in continuous recognition pipelines, and that classification performance is relatively strong. The proposed boundary refinement module provides a lightweight approach towards improving overlap consistency for application models that are prone to boundary instability without increasing network complexity, making it suitable for real-time and cloud-based applications.

Competing Interests: The authors declare that there is no conflict of interest.

Funding Information: This research did not receive grants/funds from any public or commercial agencies.

Author contribution: Author1 conceptualized defining the research objectives by developing and proposing and implementing a two-stage framework for hand gesture spotting and recognition. The author2 also conducted the process of data collection, conducting the data analysis, and conduct the experiments. Author3 and Author4 initiated the manuscript writing and contributed to revisions and modifications for preparing the final manuscript. Author2 also revised the paper based on the research requirements and provided insights throughout the preparation of the manuscript.

Informed Consent: Not Applicable

References:

Benitez-Garcia, G., Haris, M., Tsuda, Y., & Ukita, N. (2022). Continuous finger gesture spotting and recognition based on similarities between start and end frames. *IEEE Transactions on Intelligent Transportation Systems, 23*(1), 296–307. https://doi.org/10.1109/TITS.2020.3010306.

Benitez-Garcia, G., Haris, M., Tsuda, Y., & Ukita, N. (2020). Finger gesture spotting from long sequences based on multi-stream recurrent neural networks. *Sensors, 20*(2), 528. https://doi.org/10.3390/s20020528.

Chen, D., Chen, M., Wu, P., et al. (2025). Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition. *Scientific Reports, 15*, 4982. https://doi.org/10.1038/s41598-025-87752-8.

Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., & Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13359–13368). https://doi.org/10.1109/ICCV48922.2021.01311.

Chi, S., Chi, H.-G., Huang, Q., & Ramani, K. (2025). InfoGCN++: Learning representation by predicting the future for online skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*(1), 514–528. https://doi.org/10.1109/TPAMI.2024.3466212

Cui, H., Huang, R., Zhang, R., & Hayama, T. (2025). DSTSA-GCN: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. *Neurocomputing, 637*, 130066. https://doi.org/10.1016/j.neucom.2025.130066.

Emporio, M., Ghasemaghaei, A., & Laviola, J. J. (2025). Continuous hand gesture recognition: Benchmarks and methods. *Computer Vision and Image Understanding, 259*, 104435, https://doi.org/10.1016/j.cviu.2025.104435.

Gao, Y., Duan, X., & Dai, Q. (2025). Skeleton-based action recognition using graph convolutional network with pose correction and channel topology refinement. *Computers, Materials & Continua, 83*(1), 701–718. https://doi.org/10.32604/cmc.2025.060137.

Li, D., Tang, Y., Zhang, Z., & Zhang, W. (2023). Cross-stream contrastive learning for self-supervised skeleton-based action recognition. *Image and Vision Computing, 135*, 104689. https://doi.org/10.1016/j.imavis.2023.104689.

Liu, D., Zhang, L., & Wu, Y. (2022). LD-ConGR: A large RGB-D video dataset for long-distance continuous gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3294–3302). https://doi.org/10.1109/CVPR52688.2022.00330.

Liu, J., Wang, X., Wang, C., Gao, Y., & Liu, M. (2024). Temporal decoupling graph convolutional network for skeleton-based gesture recognition. *IEEE Transactions on Multimedia, 26*, 811–824. https://doi.org/10.1109/TMM.2023.3271811

Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., & Chiaberge, M. (2022). Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition, 124*, 108487. https://doi.org/10.1016/j.patcog.2021.108487.

Narayan, S., Mazumdar, A. P., & Vipparthi, S. K. (2023). SBI-DHGR: Skeleton-based intelligent dynamic hand gestures recognition. *Expert Systems with Applications, 232*, 120735. https://doi.org/10.1016/j.eswa.2023.120735.

Plizzari, C., Cannici, M., & Matteucci, M. (2021). Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding, 208*, 103219. https://doi.org/10.1016/j.cviu.2021.103219.

Ran, R., & Yang, W. (2025). FD-GCN: Feedback directed graph convolutional network for skeleton-based action recognition. *Graphical Models, 142*, 101306. https://doi.org/10.1016/j.gmod.2025.101306.

Ray, A., Raj, A., & Kolekar, M. H. (2025). Autoregressive adaptive hypergraph transformer for skeleton-based activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 938–947). https://doi.org/10.1109/WACV61041.2025.00938.

Ren, B., Liu, M., Ding, R., & Liu, H. (2024). A survey on 3D skeleton-based action recognition using learning method. *Cyborg and Bionic Systems, 5*, 0100. https://doi.org/10.34133/cbsystems.0100.

Wang, P., Wen, J., Si, C., Qian, Y., & Wang, L. (2022). Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. *IEEE Transactions on Image Processing, 31*, 6224–6238. https://doi.org/10.1109/TIP.2022.3207577.

Xin, W., Liu, R., Yi, L., & Chen, Y. (2023). Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing, 537*, 289–304. https://doi.org/10.1016/j.neucom.2023.03.001.

Yang, H., Yan, D., Zhang, L., Li, D., Sun, Y., You, S., & Maybank, S. J. (2022). Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing, 31*, 164–175. https://doi.org/10.1109/TIP.2021.3129117.

Zhou, Y., Cheng, Z.-Q., Li, C., Fang, Y., Geng, Y., Xie, X., & Keuper, M. (2022). Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*. https://doi.org/10.48550/arXiv.2211.09590.

Zhu, Q., & Deng, H. (2023). Spatial adaptive graph convolutional network for skeleton-based action recognition. *Applied Intelligence, 53*, 17796–17808. https://doi.org/10.1007/s10489-022-04442-y.

D V Srihari Babu
Department of Electronics and Communication Engineering,
G pulla Reddy Engineering College,
Kurnool,
Andhra Pradesh,
India
srihari2k1@gmail.com
ORCID 0009-0007-6238-6040

Bakkala Santha Kumar
Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation,
Vaddeswaram,
Andhra Pradesh
India
santhakumar.bakkala@gmail.com
ORCID 0009-0007-4739-0498

Y Narasimha Reddy
Department of CSE,
St John's College of Engineering and Technology,
Kurnool,
Andhra Pradesh
India
narasimhareddy.mtech@gmail.com
ORCID 0000-0003-0829-7323

M Nagaraju
Department of CSE (Artificial Intelligence and Machine Learning),
Institute of Aeronautical Engineering,
Hyderabad,
Telangana,
India
nagaraj.mavilla29@gmail.com
ORCID 0000-0001-8898-1090