CTU Journal of Innovation and Sustainable Development
ISSN 2615-9422 | e-ISSN 2815-5602

DOI:10.22144/ctujoisd.2023.045

Exploring MediaPipe optimization strategies for real-time sign language recognition

Nguyen Phuoc Thanh*, Nguyen Thanh Hoang, Hoang Ngoc Xuan Nguyen, Phan Huynh Thanh Binh, Vu Hoang Son Hai, and Huynh Hieu Nhan

Artificial Intelligence, FPT University, Viet Nam

**Corresponding author (ngpthanh15@gmail.com)*

Article info.

Received 31 Jul 2023
Revised 18 Sep 2023
Accepted 2 Oct 2023

Keywords

LSTM, MediaPipe, How2Sign, Indian Sign Language, ISL

ABSTRACT

The present study meticulously investigates optimization strategies for real-time sign language recognition (SLR) employing the MediaPipe framework. We introduce an innovative multi-modal methodology, amalgamating four distinct Long Short-Term Memory (LSTM) models dedicated to processing skeletal coordinates ascertained from the MediaPipe framework. Rigorous evaluations were executed on esteemed sign language datasets. Empirical findings underscore that the multi-modal approach significantly elevates the accuracy of the SLR model while preserving its real-time capabilities. In comparative analyses with prevalent MediaPipe-based models, our multi-modal strategy consistently manifested superior performance metrics. A distinguishing characteristic of this approach is its inherent adaptability, facilitating modifications within the LSTM layers, rendering it apt for a myriad of challenges and data typologies. Integrating the MediaPipe framework with real-time SLR markedly amplifies recognition precision, signifying a pivotal advancement in the discipline.

1. INTRODUCTION

Sign language, a vital communication mechanism for the deaf and hard-of-hearing community, relies on intricate hand gestures, body movements, and facial expressions. Technological advancements have intensified the demand for automated sign language recognition (SLR) systems to translate these gestures into comprehensible text or speech (Shi et al., 2020). While numerous methodologies have been proposed in the SLR domain, the real-time recognition aspect remains challenging because of the subtle nuances of sign language.

Driven by the MediaPipe (Lugaresi et al., 2019) framework, our research introduces a groundbreaking multi-modal SLR methodology. Central to our approach is integrating four distinct Long Short-Term Memory (LSTM) (Staudemeyer & Morris, 2019) models hinged on skeleton coordinates extracted from MediaPipe (Lugaresi et al., 2019), capturing the depth and dynamism of sign language.

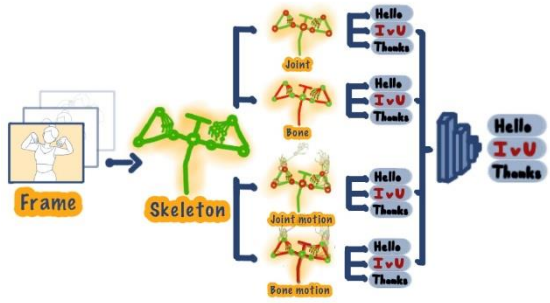

Figure 1. Proposed multi-modal SLR model using skeleton posing to enhance recognition

In urban centers like Ho Chi Minh City, the bustling heart of Vietnam, there are many stories of

individuals relying on sign language daily. Amidst the city's cacophonous streets and alleyways, the silent world of these individuals intersects with ours. A refined sign language recognition system, such as the one proposed, can facilitate more seamless interactions in such dynamic environments, creating more inclusive urban societies.

Our choice of the How2Sign dataset, as presented by (Duarte et al., 2021), underpins our methodology. This dataset's reputation in the SLR community renders it an ideal benchmark to validate the efficacy of our multi-modal approach. Our venture extends beyond merely demonstrating increased accuracy; we posit our multi-modal approach as a versatile and efficient solution for real-time SLR challenges.

While foundational, our proposed LSTMs (Staudemeyer & Morris, 2019) are inherently adaptable. The emphasis is on the overarching multi-modality, allowing the LSTMs (Staudemeyer & Morris, 2019) to be tailored to various datasets and problems rather than being a monolithic solution. This adaptability ensures that our framework remains relevant across diverse challenges.

We further align our work with state-of-the-art models rooted in the MediaPipe (Lugaresi et al., 2019) framework, specifically those highlighted by (Velmathi & Goyal, 2023). Our comparative analysis is marked by meticulous customizations to the LSTM networks (Staudemeyer & Morris, 2019), ensuring they resonate with the challenges these models address. While retaining our approach's core principles, this alignment guarantees an equitable evaluation.

The subsequent sections will offer a more granular exploration of our methodologies, evaluations, and discussions, setting the stage for continued innovations in this pivotal domain.

2. MATERIALS AND METHOD

2.1. Datasets Overview

Our research uses two cornerstone datasets: the How2Sign dataset (Duarte et al., 2021) and an Indian Sign Language dataset.

The How2Sign dataset, as introduced by (Duarte et al., 2021), has garnered attention for its pioneering attributes, emerging as the foremost expansive, multi-modal, and multi-view continuous dataset dedicated to American Sign Language (ASL). Encompassing a staggering 80 hours of sign language videos, it seamlessly integrates various modalities such as speech, meticulously curated English transcripts, and depth, setting a new standard in sign language datasets. One of its standout features is a specialized subset endowed with intricate 3D pose estimations, a feat achieved by the advanced Panoptic studio. The topics it encapsulates range from everyday subjects like "Cars and Other Vehicles" to niche themes like "Sports and Fitness," providing a comprehensive and varied learning environment. This diversity ensures a robust and holistic training paradigm, enhancing the generalizability of models trained on it. Delving deeper into our experimental design, we harness the potential of the Green Screen RGB clips from this dataset. This subset, comprising 35,191 clips meticulously extracted from 2,456 videos, presents a pragmatic yet challenging testbed for our model evaluations, simulating real-world conditions with its intricate gestures and nuances. Table 2 further explains the detailed division and categorization of the dataset.

Table 2. How2Sign Dataset Statistics

Subsets	Words	Sentences	Clips
Training	15,686	31,128	31,128
Validation	3,218	1,741	1,741
Testing	3,670	2,322	2,322
Total		35,191	35,191

To enhance the comprehensiveness of the Indian Sign Language dataset (Velmathi & Goyal, 2023), which boasts an extensive collection of 1200 photographs for each object within the dataset, our study incorporates a diverse range of elements. This dataset, crafted with meticulous attention, specifically encompasses the digits 1 to 9 and the complete alphabet from A to Z, resulting in a comprehensive representation of sign language gestures. This extensive scope ensures that our multi-modal models can be systematically compared and benchmarked against the real-time sign language recognition model developed by Velmathi and Goyal in 2023. Throughout the process of conducting these comparative analyses, we strategically adapted our LSTM network, as initially proposed by Staudemeyer & Morris in 2019, to effectively align with the intricacies and challenges posed by the model presented by Velmathi & Goyal while maintaining the core principles and methodology of our approach.

The judicious selection and combination of these datasets in our research provide a foundation for robust evaluation and benchmarking. By harnessing

the depth of How2Sign (Duarte et al., 2021) and the specificity of the Indian Sign Language dataset, we strive to push the boundaries of AI-based sign language recognition, ensuring our results are comprehensive and contextually relevant.

2.2. Sign Recognition Techniques

Sign language is a beacon of effective communication for those with hearing and speech impairments, which is pivotal in fostering connections and even advancing cognitive development in specific contexts (Emmorey, 2001). As essential as it is, the real-time deciphering of sign language and the demand for impeccable accuracy is an intricate challenge (Dardas & Georganas, 2011). While traditional strategies like Convolutional Neural Networks (CNNs) (Huang et al., 2018) have showcased promise in static image recognition, their computational demands escalate when real-time analysis comes into play. Our research aims to break this impasse, offering a rejuvenated perspective on the problem. Our solution revolves around skeleton pose estimation, a method that can transcend the confines of previous strategies, as visually portrayed in Figure 2.

Historically, sign language recognition has relied heavily on techniques steeped in CNNs (Huang et al., 2018), processing gestures extracted from static images. The results, undeniably potent, have not been without their pitfalls. Every frame in a video stream funnels a significant computational toll, stretching resources thin. This computational heftiness translates to a bottleneck in real-time gesture recognition, introducing compromises in accuracy to maintain timeliness. Moreover, sign language's richness extends beyond mere hand gestures; facial cues and body postures add layers of meaning and, simultaneously, layers of computational demands.

Venturing off the trodden path, our proposal emphasizes skeleton pose estimation, zeroing in on vital joints and landmarks rather than canvassing the entire frame. This refined approach bears a host of benefits. Primarily, it slashes computational requirements by narrowing the focus to skeletal configurations. With less data to wrangle, computational speeds surge, paving the way for accurate real-time recognition. An added boon is our model's indifference to video resolution; its essence lies in interpreting joints and landmarks, not pixel densities. Furthermore, we have designed our model to exhibit resilience, gracefully navigating challenges like fluctuating lighting conditions, obstructions, and varied backgrounds.

The finesse of our skeleton's estimation owes much to MediaPipe (Lugaresi et al., 2019). Esteemed in computer vision for its blend of rapidity and precision, MediaPipe's deep learning modules (Lugaresi et al., 2019) are fine-tuned for real-time tasks. Its prowess shines brightly in landmark detection, a facet we exploit to its fullest. In deciphering sign language gestures, we zero in on a meticulously curated set: 34 posture points, a dense grid of 468 facial landmarks, and 21 points each for the left and right hands. This comprehensive constellation guarantees that no subtle or overt nuance escapes our notice. The fusion of MediaPipe's impeccable landmark detection with our methodology sets the stage for a groundbreaking sign language recognition system.

Our foray into sign language recognition, anchored by the tenets of skeleton pose estimation and supercharged with MediaPipe's proficiency (Lugaresi et al., 2019), heralds a fresh era in this domain. Its ramifications extend beyond academic intrigue, holding tangible promise in areas like assistive technologies for hearing-impaired and novel educational tools.

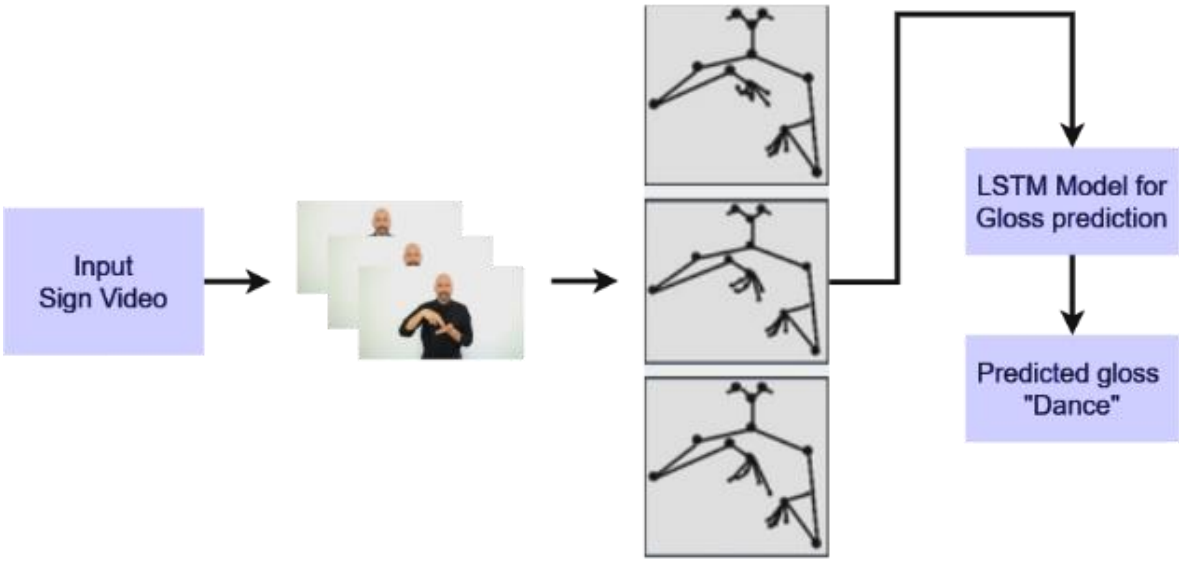

Figure 2. Real-Time Sign Language Recognition via Skeleton Pose Estimation

2.3. Multi-Modal Multi-Stream Approach

In the realm of sign language recognition, real-time performance is a cardinal necessity. Building upon the principles shown in Figure 3, our methodology brings forth a multi-stream approach. This approach is meticulously tailored to confront the complexities of discerning sign language gestures, primarily when solely relying on skeleton coordinates within short bursts of 30 frames. The crux of the challenge lies in the inherent similarity of skeletal movements in continuous motion-based signs. Our answer to this is an ensemble of models, each homing in on different skeletal aspects, functioning in tandem to yield optimal real-time results.

Joint Sign Language Recognition Model: Our journey begins with the foundational concept of interpreting sign languages directly via skeletal joint coordinates, represented as:

$$\mathbf{J}(\mathbf{x_i^t}, \mathbf{y_i^t}, \mathbf{z_i^t})$$

In this representation, $\mathbf{x_i}$ and $\mathbf{y_i}$ pinpoint a skeletal joint's position within a singular video frame. Concurrently, $\mathbf{z_i}$ conveys our confidence in that specific joint detection, with $\mathbf{t}$ emphasizing the temporal progression across frames. For example, inputting a video frame by frame will output a series of coordinates of the joints from Mediapipe.

Bone Sign Language Recognition Model: As we progress, we introduce a model underpinned by bone motion. Recognizing the dynamism of bone interactions during sign execution, vectors, formulated as:

$$\mathbf{B}(\mathbf{x_j^t} - \mathbf{x_i^t}, \mathbf{y_j^t} - \mathbf{y_i^t}, \mathbf{z_i^t})$$

Become pivotal. These vectors encapsulate the nuances of bone position shifts, serving as a beacon to capture the unique movement blueprints tied to diverse sign language expressions.

Joint Motion Sign Language Recognition Model: Pivoting to a more holistic view, our third model amalgamates a heatmap representation, offering a temporal vista of how skeletal joint coordinates transition across frames. This visualization approach is potent in differentiating skeletal setups that may otherwise seem identical. To augment this, we integrate joint motion insights by assessing consecutive frame disparities, captured as:

$$\mathbf{JM}(\mathbf{x_i^{t+1}} - \mathbf{x_i^t}, \mathbf{y_i^{t+1}} - \mathbf{y_i^t}, \mathbf{z_i^t})$$

offering a granular lens into skeletal transitions. For example, when inserted in 31 frames, it will get 31 skeleton coordinates. From these 31 Joint coordinates, we will rely on the formula and calculate 30 Joint Motions.

Bone Motion Sign Language Recognition Model: Our model ensemble's culmination is deeply entrenched in the temporal domain. It delves into the granularities of frame-to-frame skeletal shifts, offering a vantage point from where even minute motion subtleties are perceptible. The key lies in assessing bone vector transitions between frames, represented succinctly as:

$$\mathbf{BM}(\mathbf{x_j^{t+1}} - \mathbf{x_i^{t+1}} - \mathbf{x_j^t} + \mathbf{x_i^t}, \mathbf{x_j^{t+1}} - \mathbf{x_i^{t+1}} - \mathbf{x_j^t} + \mathbf{x_i^t}, \mathbf{z_i^t})$$

The training regiment for these models mandates independence, ensuring each becomes adept at recognizing actions from the primordial landmark points. As the essence of bone and joint motion lies in frame sequences, our blueprint embraces the inaugural model's structure, evaluating 30 frames consecutively.

Fusing insights, our ensemble technique amalgamates predictions from each model, yielding the cohesive sign recognition verdict:

$$\mathbf{q} = \mathbf{J}.\propto_1 + \mathbf{B}.\propto_2 + \mathbf{JM}.\propto_3 + \mathbf{BM}.\propto_4$$

Herein, J, B, JM, and BM symbolize predictions from the individual models: Joint, Bone, Joint Motion, and Bone Motion. The weightage coefficients $\propto_1$, $\propto_2$, $\propto_3$,and $\propto_4$ dictate the relative significance of each model's insights.

With our multi-stream ensemble, the objective transcends mere accuracy enhancement. We sculpt a resilient, comprehensive framework for real-time sign language recognition, amalgamating spatial nuances, bone and joint dynamics, and temporal intricacies into a singular, potent methodology.

2.4. Neural Network

The neural network's architecture is paramount in our journey through the multi-modal, multi-stream approach. This section introduces the foundational LSTM model (Staudemeyer & Morris, 2019) structure for clarity and replication. It is essential to understand that while this LSTM blueprint serves as a minimalist yet efficient starting point, its configuration and hyperparameters might necessitate adjustments based on specific datasets and the intricacies of problems at hand.

Long Short-Term Memory (LSTM) (Staudemeyer & Morris, 2019) is a specialized variant of Recurrent Neural Networks (RNN) (Sofianos et al., 2021), meticulously crafted to counter the vanishing gradient problem inherent in traditional RNNs (Staudemeyer & Morris, 2019). The heart of LSTM (Staudemeyer & Morris, 2019) lies in its distinctive unit, encompassing a cell and three gates: an input gate, an output gate, and a forget gate. This unique structure enables the cell to retain values across varied time intervals, with the gates deftly regulating the information flow in and out of the cell.

The dense layer, often considered fully connected, functions as the melting pot of abstract representations. It is achieved by intricately connecting neurons to every preceding layer's neuron. In parallel, we integrate the Dropout layer to ensure model robustness and deter over-fitting (Srivastava et al., 2014). This regularization technique periodically and randomly deactivates specific input units during the training phase, ensuring the model does not develop an excessive dependency on particular training data patterns.

Activation functions are the neural network's linchpin, catalyzing non-linearity and driving intricate mappings between inputs and outputs. The

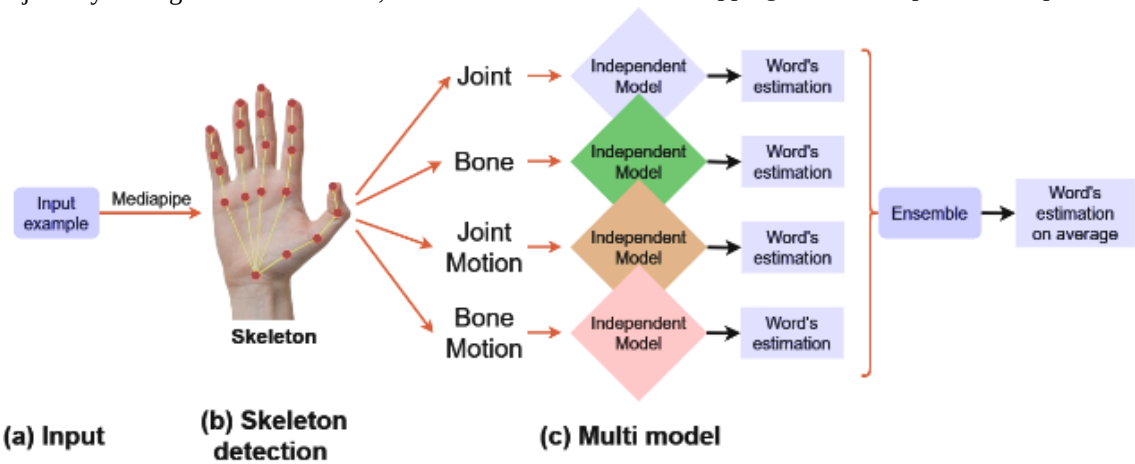

Figure 3. Illustration of the model pipeline: Process visualization

Rectified Linear Unit (ReLU) (Agarap, 2019) is an elemental piece-wise linear function, which, if fed a positive input, echoes it and, for non-positive inputs, returns zero. Its widespread adoption stems from the ease of training models that use it, often yielding superior performance. Conclusively, the neural network leverages the soft-max activation function (Pearce et al., 2021), which metamorphoses raw neural outputs into a structured probability vector, a

delineated probability distribution across input categories. Mathematically, the soft-max activation function is represented as:

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{j-1}^{N} e^{x_j}}$$

Here, x_i embodies the input value of the i^{th} element within the input vector, with N symbolizing the count of input vector elements.

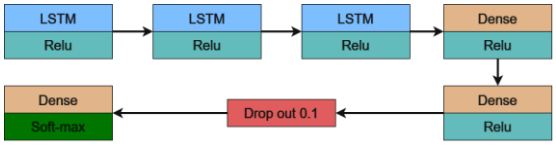

Figure 4. Schematic of the Neural Network Framework

In summary, our proposed architecture pivots around a foundational LSTM structure adeptly designed to interpret 30 NumPy arrays and yield a probability array for prospective actions. While this framework forms the bedrock of our multi-modal ensemble, it is paramount to acknowledge its malleability. Real-world applications may necessitate tailored refinements, ensuring the architecture resonates harmoniously with the idiosyncrasies of distinct datasets and their myriad challenges. As the world of sign language recognition forges ahead, such adaptable blueprints will invariably dictate the trajectory of advancements.

3. RESULTS AND DISCUSSION

Embarking on this journey with the comprehensive How2Sign dataset (Duarte et al., 2021), we trained and meticulously evaluated four distinct sign language recognition models: the joint model, bone model, joint motion model, and bone motion model Figure 3. Beyond merely understanding their performance, our focus intensified as we customized our Multi-modal application for a deeper comparative analysis against an Indian model (Velmathi & Goyal, 2023) using the Indian Sign Language dataset. In ensuring a fair comparison, while we tailored our LSTM layer to resonate with the dataset and challenges targeted by the (Velmathi & Goyal, 2023) model, the essence and principles of our multi-modal approach remained unaltered. This meticulous approach safeguards our research from biases and underscores the integrity of our results. As we move forward, this section also sets the stage to discuss prospects, culminating in unveiling our FPTs2l application Figure 5, a manifestation of our scientific endeavors to revolutionize human-computer interactions for those with hearing impairments.

3.1. How2Sign Dataset Evaluation

The How2Sign dataset, as delineated by (Duarte et al., 2021), serves as a veritable treasure trove for sign language researchers, a sentiment manifested in Table 2. Its vast repository, encompassing varied sign language expressions across multiple practitioners against a standardized green screen backdrop, is pivotal for precise pose approximations, consequently bolstering the accuracy and reliability of sign language recognition systems.

An elaborate dive into our model training sequence provides insights into its sophistication. Initially, we engaged with a comprehensive collection of 35,191 clips, each correlating to a unique sentence. This vastness translates to 35,191 distinct actions, setting the groundwork for subsequent steps. These clips were then dissected frame by frame, and each became a substrate for MediaPipe, a state-of-the-art tool tailored for extracting joint coordinates. After this extraction, two primary processes were employed: the calculation of Bone vector coordinates and the intricate computation of Joint Motions. While the former rests on the skeleton coordinates on individual frames, the latter draws from comparing coordinates between consecutive frames. An implication of this methodology is the emergence of 31 NumPy arrays for the Joint and Bone models from a video spanning 31 frames and only 30 NumPy arrays for both the Joint Motion and Bone Motion models because of the frame-to-frame comparisons given the negligible impact of a single frame on action interpretation, a solitary NumPy array was dismissed, ensuring alignment across models and diminishing any induced bias. Crucially, each model was architectured with four LSTM layers, adeptly equipped to process sets of 30 consecutive NumPy arrays.

Having distilled the dataset into its most refined form, our next endeavor was architecting a Neural LSTM network. The design of this network was congruent with the foundational neural network posited for our four models. The training regimen for this model was fortified with the Adam optimization technique, and the categorical-cross-entropy loss function was employed to minimize discrepancies. The meticulous dataset segmentation into training, validation, and testing sets is worth noting, a decision vindicated by the dataset's

authors, who posited this division as the most optimal for ensuring an equitable distribution.

Advancing to the next echelon, the LSTM network's design was symbiotic with a rudimentary neural network mapped out for the quartet of models. This network was subjected to the Adam optimization technique, and errors were curtailed by employing the 'categorical-cross-entropy' loss function. The dataset's trifurcation into training, validation, and testing segments heeds the guidelines advocated by the authors, asserting an optimal and balanced distribution. Each LSTM operation ingests 30 meticulously computed NumPy arrays as input, rendering an array illuminating probabilities linked to the pre-defined 35,191 actions.

The crescendo of our methodology lies in the Multi-Modal approach, wherein $\propto_n$ parameters are initialized at 0.25 for all constituent models, mirroring the multivariate regression model's orientation. The Multi-Modal's genius rests in synthesizing four probability arrays, ingeniously adjusting them using the predetermined alpha parameters from the multivariate regression model. The resultant array, distilled from the harmonization of the four arrays, prompts the selection of an action boasting the apex probability.

In culmination, as evident in Table 3, our models, once trained on the How2Sign dataset, yielded fascinating insights. The Joint model emerged as the front-runner in accuracy among the four individual models. This reaffirms the longstanding belief in.

Table 3. Performance of Different Streams

Models	Precision	Accuracy	F1-score
Joint	0,681	0,67	0,653
Bone	0,678	0,668	0,651
Joint Motion	0,666	0,656	0,639
Bone Motion	0,663	0,652	0,635
Multi-Modal	0,684	0,673	0,656

Sign language recognition frequently capitalizes on joint coordinates, owing to their ability to precisely encapsulate the essence of gestures, reflecting the intricate movements and postures of the signer. A closer look at our model performances provides an illuminating perspective on this. Specifically, the Joint model, with an accuracy of 0.67, outstrips the Bone model at 0.668, the Joint Motion model at 0.656, and the Bone Motion model, which clocks in at 0.652. The superiority of the Joint model in capturing gesture subtleties offers an evidential backdrop as to why joint coordinates are often the go-to features in many sign language recognition systems.

Nevertheless, transcending these individual performances, our Multi-Modal method, an amalgamation of the unique strengths of these models, achieves an accuracy of 0.673. This pinnacle of performance attests to the efficacy of combining diverse skeletal posing techniques. It suggests that integrating Bone, Joint Motion, and Bone Motion models can enhance the robustness of sign language recognition systems.

Pitted against the renowned Indian Sign Language Recognition (ISLR) model formulated by (Velmathi & Goyal, 2023), our Multi-Modal approach held its ground, bringing to the fore its distinctive merits and hinting at scopes for further fine-tuning and refinement.

While the inherent efficacy of the Joint model underlines the time-tested value of joint coordinates in sign language recognition, our multi-modal approach's holistic, composite nature truly shines. As we progress, such multi-faceted models have the potential to revolutionize assistive technology, making it more adaptive and intuitive in the domain of human-computer interaction.

3.2. Comparative Analysis

The nexus between MediaPipe (Lugaresi et al., 2019) and LSTM (Staudemeyer & Morris, 2019) for sign language recognition is still being discovered. Nevertheless, our presented multi-modal approach Figure 3, unveils a novel paradigm. Our approach achieves a heightened recognition accuracy by leveraging skeletal frames harnessed via MediaPipe and additional modalities.

The intricate dynamic between the MediaPipe framework (Lugaresi et al., 2019) and LSTM (Staudemeyer & Morris, 2019) in the context of sign language recognition remains a fertile ground for exploration. In this realm, our multi-modal methodology, illustrated in Figure 3, offers a novel trajectory. By harnessing skeletal information sourced through MediaPipe and supplementing it with various modalities, we achieve a marked enhancement in recognition accuracy.

To rigorously scrutinize the merits of our approach, we juxtaposed it with the esteemed Indian Sign Language Recognition (ISLR) model, as conceived by Velmathi & Goyal (2023).. This comparison transcends mere statistical benchmarking, providing

a panoramic view of our model's unique attributes and illuminating avenues for further evolution.

The Indian Sign Language (ISL) dataset serves as a foundational pillar in our research, encapsulating a pictorial lexicon of English alphabets from A to Z. Each letter is represented by 1,200 individual images that intricately capture the nuanced semantics inherent to each sign. Through the adeptness of the MediaPipe framework, we extracted essential skeletal data, precisely the critical joint coordinates, which were then numerically transformed and stored in NumPy arrays for swift processing. While our methodological approach broadly resonates with the protocols delineated by (Velmathi & Goyal, 2023), explicitly employing the categorical cross-entropy loss function, a distinct contrast is evident when juxtaposed with the How2Sign dataset's modus operandi. Whereas our ISL evaluations are anchored in individual images, the How2Sign paradigm harnesses dynamic, clip-centric sequences, showcasing the flexibility of our multi-faceted model that seamlessly integrates with static and sequential data sources.

A comprehensive perusal of Table 4 not only reveals the performance metrics of our model concerning the ISLR model but also serves as an emblematic testament to our model's capabilities, the challenges encountered, and the potential vistas it opens in leveraging the MediaPipe framework for sign language recognition. The Table accentuates our approach's relative prowess and prospective enhancements, making a compelling case for its practical applicability in real-world scenarios.

Table 4. Performance of Different Models

Models	Accuracy	F1-score	Average Execution Times
ISLR	0,855	0,847	0,07s
Multi-Modal	0,858	0,85	0,3s

At the nucleus of any recognition system lies its accuracy. Our research underscores this statement's significance by demonstrating our multimodal approach's superior accuracy. Achieving an accuracy rate of 0.858, our model presents a modest but statistically significant advantage over the ISLR model's 0.855 (Velmathi & Goyal, 2023). This slight edge suggests our model's enhanced proficiency in discerning gestures in Indian Sign Language.

Furthermore, if realized in an authentic system design, the capability to run our model across four independent reading streams can equalize or even offset the time execution disparities. This hints at the scalability and adaptability of our approach in diverse system architectures. Our model does not merely match the standards. It often exceeds them, displaying competitive metrics that further accentuate its robustness.

However, every coin has two sides. Our multimodal approach exhibits an execution time of 0.3s, slightly increasing when juxtaposed with the ISLR model's 0.07s. Nevertheless, it is crucial to illuminate that this latency, albeit higher, still firmly remains within the realms of real-time execution constraints. To offer perspective, given a standard of 30 frames per second (FPS) for real-time video, our model's latency equates to processing roughly nine frames a performance metric that aligns with instantaneous applications.

While our multimodal approach entails a moderate trade-off concerning execution time, its merits, primarily in terms of accuracy, position it as a formidable contender in Indian Sign Language recognition. With continuous refinements, it might eclipse the renowned ISLR model in future applications.

3.3. Discussion

Accuracy is undeniably paramount in assessing the effectiveness of sign language recognition models. It provides an insightful ratio, elucidating the number of correct predictions concerning all predictions rendered. This metric's significance in sign language recognition is unparalleled, especially compared to other performance metrics, such as recall or precision. The reasoning lies in the inherent nature of sign language communication. For instance, while signing a phrase like "How are you?", there are often minuscule pauses between the words "How," "Are," and "you." These fleeting moments of stillness are as essential to the integrity of the message as the signs themselves. In another example, consider a signer articulating a more complex statement like "Although it is raining, I would like to go out." The pauses here might be slightly elongated, especially before the contrasting phrase "I would like to go out." Focusing only on recall could risk amplifying the detection of active signs, but might marginalize these integral pauses. Similarly, an overemphasis on precision could inadvertently filter out these gaps, presuming them as noise or irrelevant data. Moreover, in practical

scenarios, signers might occasionally incorporate non-linguistic or cultural gestures, such as nodding for affirmation or shaking the head for negation. A model fixated solely on recall or precision could misinterpret these as signs or miss them entirely. In stark contrast, accuracy stands out as it holistically encompasses both the articulated signs and the vital intervals between them, ensuring that a model captures signs and effectively distinguishes them from other peripheral movements or pauses.

Employing categorical entropy loss is commonplace in multi-class classification scenarios, fitting snugly within the purview of sign language recognition. By minimizing this entropy loss, we endeavor to enhance accuracy, effectively curtailing incorrect predictions. This loss function can be a beacon, illuminating underlying data inconsistencies like imbalances or noise, paving the path for corrective strategies, such as data augmentation or regularization.

Our method, pivoting on the Long Short-Term Memory (LSTM) neural networks (Staudemeyer & Morris, 2019), marks a significant departure from conventional approaches. While earlier works revolved around recognizing individual gestures, often limited to distinct letters or words, our methodology embraces the challenge of deciphering sequences to yield coherent phrases or sentences. However, a discernible challenge surfaces when processing individual letters or standalone words. For instance, the sign for the letter "A" might bear semblance to a thumbs-up gesture, mainly if observed in isolation without the context of an ongoing conversation. Such overlaps can muddy the waters, causing our model to occasionally falter due to these ambiguities and other gestures unrelated to sign language.

Leveraging the How2Sign dataset (Duarte et al., 2021) as our testing bed, we found the Joint model, which concentrates on joint coordinates, consistently outshining its counterparts in terms of accuracy, a testament to the prevailing preference of employing joint coordinates in sign language recognition (Velmathi & Goyal, 2023). Nevertheless, the crown undoubtedly rests with our Multi-Modal model, achieving unparalleled accuracy and accentuating the advantage of integrating multiple modalities.

Looking beyond the immediate domain of sign language recognition, the implications of our multi-modal method are manifold. Though our current model shows laudable speed and efficacy in capturing real-time gestures, its adeptness in impeccable sign language translation warrants further refinement. Nevertheless, the potential applications are vast: from gesture-based smart device operations, identifying suspicious activities or gang confrontations through specific signs to preemptively flagging occupational hazards. Our model could be the linchpin for numerous solutions, transcending the realm of mere sign language recognition.

A vital facet of our research was the conception and development of the FPTs2l application, a tangible manifestation of our approach's potential to transition from theory to praxis. This application was sculpted to test the real-time capabilities of our multi-modal methodology. Our preliminary observations were promising. The application adeptly captured gestures via camera feed, reflecting the quick responsiveness we theorized.

However, it is imperative to acknowledge its limitations. While the application detected broader gestures with commendable accuracy, it grappled with identifying nuanced gestures, particularly those representing singular words or letters. This could be attributed to the intricate resemblance between certain sign gestures and other commonplace actions. For instance, the sign for the word "water" in Vietnamese Sign Language might closely mimic a frequently adopted hand motion by users not necessarily signifying "water". These overlapping gestures present a conundrum, occasionally causing our application to need to be more accurate.

Furthermore, we used a dataset of 30 vocabulary items from online videos showcasing the most prevalent vocabulary in Vietnamese. Given the restricted dataset, the model's proficiency in comprehensive real-time translation for diverse conversations remains to be seen.

Nevertheless, it is essential to underscore the broader ramifications of our work. The FPTs2l application serves as a prototype, a forerunner to potential future applications that can span a spectrum of use cases. Think of enhanced models tailored for academic integrity by detecting suspicious gestures during online examinations. Alternatively, envision a system embedded in smart homes, translating gestures into commands for household devices. On a more ambitious note, this could morph into a security apparatus, discerning hostile intents through specific hand signs in crowded areas. The horizon is vast, from potential

occupational safety applications to conflict detection.

In conclusion, while our multi-modal approach and the FPTs2l application promise an exciting future for gesture recognition and real-time translations, we acknowledge the journey ahead with refinements, expansions, and practical testing on diverse datasets to realize its potential.

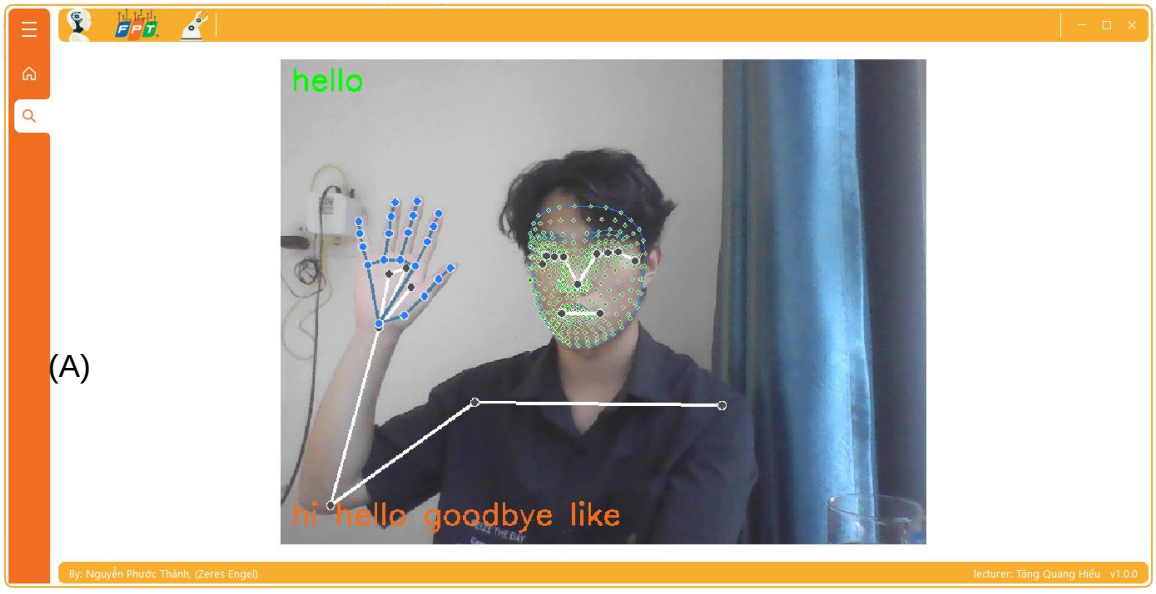

Figure 5. Screenshot of the FPTs2l Application in Action

This application, built to evaluate the real-time capabilities of our approach, successfully processes video feed and identifies gestures with a commendable frame rate of 2 FPS. Despite its efficacy, there are occasional misidentifications, emphasizing areas for improvement.

4. CONCLUSION

4.1. Significance and Implications

In this exploration, we delved into the realm of real-time sign language recognition by leveraging the capabilities of the MediaPipe framework (Lugaresi et al., 2019). We introduced a multi-modal approach that synergistically combined four LSTM models (Staudemeyer & Morris, 2019), trained on skeletal coordinates extracted using MediaPipe. This fusion capitalized on the strengths of each model, enhancing the overall recognition accuracy.

While traditional methods often get bogged down with intricate preprocessing and feature extraction, our approach, rooted in the MediaPipe framework, bypassed these complexities to offer real-time performance. This speed, combined with augmented accuracy through our multi-modal approach, establishes the potential for our model as an efficient tool for facilitating communication for the deaf and hard-of-hearing community.

However, it is noteworthy that while our method advanced in real-time processing and reduced the need for extensive preprocessing, there remains room for refining accuracy. Certain subtle or intricate gestures pose challenges. Nevertheless, this limitation offers a direction for future endeavors, suggesting that enhancements can be made by adapting the LSTM network layers according to specific problems and datasets.

4.2. Future Horizons

Considering the potential and the challenges observed, future research efforts are primed for several exciting directions. At the forefront is the aspiration to devise a real-time translation tool that promotes uninhibited communication between deaf or hard-of-hearing individuals and those unfamiliar with sign language. Such a tool would not only facilitate interactions but also stand to democratize access to information and resources.

With the proliferation of digital content, it is urgently needed to be accessible. Our approach can be further refined to auto-generate accurate subtitles or captions for various multimedia channels, ensuring the inclusivity of content consumption.

Beyond mere recognition, the realm of sign language is vast. Sign language is not merely about

hand gestures; facial expressions, emotions, and context play pivotal roles. Future explorations could delve deeper into a comprehensive multi-modal approach that integrates these facets, pushing the boundaries of accuracy and expressiveness in sign language recognition (Emmorey, 2001).

Furthermore, in the educational landscape, there lies an untapped potential to harness sign language for creating interactive games and tools. These can help to promote sign language learning, foster broader adoption, and cultivate cultural appreciation.

ACKNOWLEDGMENT

We extend our profound gratitude to the creators of the How2Sign dataset for their commendable efforts in producing such a comprehensive and invaluable dataset on sign language. Their contribution has been pivotal in facilitating our research and corroborating the efficacy of our method. We also express our appreciation to all contributors and reviewers who have imparted their invaluable insights and feedback, further strengthening the rigor and relevance of our study.

REFERENCES

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing, 29*, 9532-9545.

Dardas, N. H., & Georganas, N. D. (2011). Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and Measurement, 60*(11), 3592-3607.

Velmathi, G., & Goyal, K. (2023). *Indian Sign Language Recognition Using Mediapipe Holistic*. arXiv preprint. https://arxiv.org/abs/2304.10256

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). *MediaPipe: A Framework for Building Perception Pipelines*. arXiv preprint. https://arxiv.org/abs/1906.08172

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks*. arXiv preprint. https://arxiv.org/abs/1909.09586

Emmorey, K. (2001). *Language, cognition, and the brain: Insights from sign language research*. Psychology Press.

Huang, J., Zhou, W., Li, H., & Li, W. (2018). Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(9), 2822-2832.

Sofianos, T., Sampieri, A., Franco, L., & Galasso, F. (2021). Space-Time-Separable Graph Convolutional Network for Pose Forecasting. *CoRR*, abs/2110.04573. https://arxiv.org/abs/2110.04573

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929-1958. http://jmlr.org/papers/v15/srivastava14a.html