International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 11, November, 2025, pp. 2949-2959
Journal homepage: http://iieta.org/journals/isi

VOGUE-Based Approach for Segmenting Movement Epenthesis in Continuous Sign Language Recognition

Thillai Sivakavi S, Minu R. I.*

Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur 603203, Chengalpattu, India

Corresponding Author Email: minur@srmist.edu.in

https://doi.org/10.18280/isi.301113

Received: 1 September 2025
Revised: 10 November 2025
Accepted: 18 November 2025
Available online: 30 November 2025

Keywords:
VOGUE-HMM, Key Curvature Maximum Points (KCMP), context tree, sign language, continuous sign language recognition

ABSTRACT

When developing a continuous sign language recognition (CSLR) system, a significant challenge lies in processing the vast number of video frames, which demands extensive time and computational resources during both the training and prediction phases. To address this, we propose an efficient and scalable methodology that integrates cluster-based key frame extraction with a VOGUE-based recognition model designed for continuous gestures. The key frame extraction strategy clusters visually similar frames to reduce redundancy while preserving only those with high semantic relevance. To further enhance recognition accuracy, we introduce the Key Curvature Maximum Point (KCMP) technique, which identifies pivotal motion points and captures essential hand trajectory changes inherent to sign language. These refined frames are subsequently used to train a VOGUE-based model that encodes spatial and temporal strokes dynamics, followed by probability distribution modeling for robust prediction. The proposed approach was evaluated using a custom-built Tamil Sign Language dataset. Performance was compared against several established baseline methods, including Dynamic Time Warping (DTW), Hidden Markov Models (HMM), and multiple Conditional Random Field (CRF) variants, as well as the VOM model. The system achieved a recognition accuracy of 86.78% and a sign error rate of 5.3%. A paired t-test confirmed that the improvements over baseline models were statistically significant ($p < 0.05$). These results demonstrate that the proposed framework provides improved efficiency and competitive accuracy, offering a promising solution for real-time CSLR applications, particularly in low-resource regional sign languages.

1. INTRODUCTION

In the past, there has been a lack of attention towards sign languages in India due to the absence of a comprehensive corpus of sign languages for Indian languages. However, following the initiatives taken post 2000, such as the release of the first corpora of Indian Sign Language by Vivekananda University in collaboration with IHRDC and CBM, the emphasis has shifted towards promoting awareness and education about the sign languages used by the deaf and hard-of-hearing community. Sign language has proven to be the most effective form of communication for the deaf population, as it allows them to express themselves without causing discomfort to others. This has underscored the importance of educating the hearing community about sign language, enabling them to understand and engage in effective communication with the deaf community. Similarly to the mother tongue of a traditional culture, sign language can be considered the mother tongue of the deaf community, characterized by its own syntax, structure, and expressive elements. However, for the broader public to understand sign language, the development of efficient translator or recognizer systems is essential. These systems would decode signs into voice or text, and the complexity of such systems would depend on whether they are static or dynamic, based on the nature of the signs being conveyed. Typically, static or isolated sign language recognition (ISR) focuses on recognizing a single sign at a time, as illustrated by the sample sequence shown in Figure 1. Since the signer pauses after each sign, most solitary sign language videos do not contain non-sensical movements when processed. However, continuous sign language recognition (CSLR) is different because a continuous sign consists of multiple isolated sign words (ISR). We need to segment the isolated words to recognize each for a continuous sign language.

The signer must use a stop sign between each isolated word to differentiate between meaningful and meaningless segments in a continuous sign [1]. Using stop signs for longer phrases or paragraphs would become tedious if one were used for each word in a sentence. Therefore, there should be a method to separate the actual and meaningless segments. "Movement In Epenthesis" (ME) frames refer to frames or sections of a video that do not make sense. Figure 1 shows the possibility of the movement-epenthesis frame where a transition occurs between the end of the first word இந்த (this) and the beginning frame for the second word ஆடை (clothing). Here, the movement of hands in between the first and third frames movement epenthesis frame. This work proposes a system that identifies and eliminates ME portions before starting any recognition. The significant contributions of this work are as follows.

- The context tree will extract the movement epenthesis frames from the continuous sentence sequence and segment the meaningful portions.
- Determine whether the suggested methods are less time-consuming than current maximum likelihood-based movement epenthesis segmentation.

To address the limitations of traditional HMM-based and CRF-based movement epenthesis segmentation, this work adopts the VOGUE (Variable Order and Gapped HMM for Unstructured Elements) model. Unlike classical HMMs, which assume fixed-order dependencies and rely on Viterbi decoding with high computational cost, VOGUE learns variable-length contextual patterns directly from the mined strokes sequences. It also models the gaps (durations) between strokes, enabling the system to naturally handle irregular and highly variable ME segments. These properties make VOGUE particularly suitable for continuous sign language, where ME durations are unpredictable and transition patterns differ across signers. In addition, VOGUE performs likelihood computation in linear time, offering a significant speed advantage compared to the cubic time complexity of traditional HMM and CRF segmentation methods.

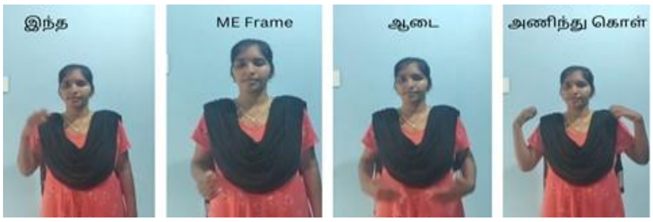

Figure 1. Sample movement epenthesis frames for a sentence in the Tamil language

The remainder of the work is structured as follows. Section 2 explains the state-of-the-art work that attempted to solve the movement epenthesis problem. Section 3 presents the proposed approach, followed by the analysis of the results in Section 4. Section 5 concludes the work.

2. RELATED WORKS

This section discusses the recognition of sign language, particularly addressing the issue of movement epenthesis. Researchers utilize various mathematical models, including the hidden Markov model (HMM), conditional random fields (CRF), dynamic temporal warping (DTW), and deep learning models, to identify signs in continuous sign language.

2.1 Movement epenthesis problem solving using Dynamic Temporal Warping (DTW)

The dynamic time-warping approach calculates the distance between two-word frames of different lengths and is the first step to find the epenthesis frame of movement [2, 3]. The main idea behind employing DTW in sign language recognition is to recognize identical sign sequences with changing speed and duration.

This is helpful when the statement is brief, but it is more challenging to match the signs when the sentence has several words. This is the leading cause of the reduction in DTW-based solutions for sign language recognition. Although comparing signs in DTW can be challenging, things become even more difficult when dealing with the movement epenthesis frames [4-6] point out that several movement epenthesis frames must be considered while developing a sign language detection system. We need a sizable corpus of ME frames, making it unable to apply the basic strategy of matching frames to corpora. Since ME frames are often present between the conclusion of the first sign and the beginning of the second sign, this strategy is the most practical. Manuel et al. [7] applied DTW in the AMBOC system for speaker verification, demonstrating its effectiveness in aligning variable length audio sequences a concept similarly applicable in gesture and sign segmentation were speed and transition variability impact accuracy.

Consider two neighboring signs, called “first” and “second”, to illustrate this mathematically. The transition movement from first to second may be calculated as $P(second|first)$. They claim that clustered models are the best at handling sentences with larger words since they have discovered that two adjacent signs will always have a near-end and a start.

The automatic separation of transition movement from real word signs is accomplished using a suggested training algorithm. The major problem identified with the Gao approach is that, irrespective of different sentences, the ME between two adjacent signs will always be the same. This is addressed by the process adopted by the study [8] using the enhanced level building algorithm. Using this strategy within the dynamic programming framework, model creation for the ME frames is intended to be avoided. The standard approaches are compared using a schematic in Figure 2.

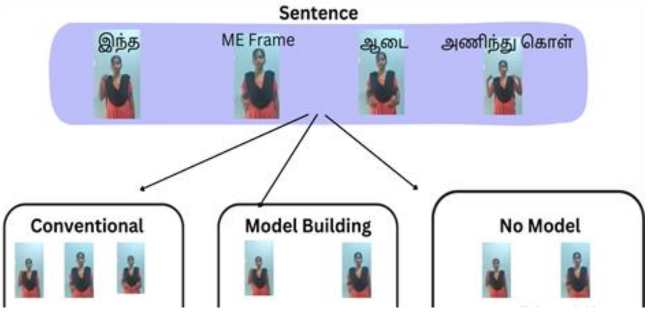

Figure 2. State-of-the-art approaches using dynamic programming to solve the problem of movement epenthesis

The ME frames could be interpreted as a sign. This is because standard methods, as shown in Figure 2, do not take them into account. When the sentence is long, it becomes practically difficult to build a model, and the no model approach aims to match the sign corpus. If a match cannot be found, it labels those frames as ME frames. The researchers focused on using Markov models to address the movement epenthesis problem because the identification of ME frames using DTW was unsatisfactory in both methods. There is also literature evidence demonstrating that the DTW approach performed better for character recognition than for sign recognition.

2.2 Movement epenthesis problem solving using Markov models

The variation of sign language over time is a crucial factor for continuous sign language recognition systems. Models capturing temporal patterns can aid in sign classification, especially for time-varying signs. The Markov model is capable of handling spatiotemporal fluctuations in a sequence, while the hidden Markov model can identify unobserved information with the help of observed data [9]. The study [8] demonstrated that modeling movement epenthesis frames significantly improved the performance of sign categorization using the HMM model. Additionally, the study [10] extended the HMM model for gesture sign recognition to recognize continuous sign language. For Korean continuous sign language, the study [11] proposed an automata-based method considering hand movements and speed as the primary study parameter. They suggested a 3-phase system: preparation, strokes, and end. Each phase represents a state in automata theory, and hand motions are classified into 18 classes using HMM [12]. The study [13] utilized the threshold model labeling method and the quick HMM algorithm to enhance the recognition system's performance, resulting in a lower error rate of 12.2%. However, due to extended dependencies, they found that the HMM-based model was not able to account for all the necessary relationships, leading to a shift in favor of the CRF method.

2.3 Epenthesis problem solving using Continuous Random Field (CRF)

The Continuous Random Field (CRF) is more flexible in associating labels with observations than Hidden Markov Models (HMM) [14]. This is because CRF focuses on directly collecting posterior probabilities [15], allowing for the creation of a single model for all sign labels. However, setting a fixed threshold for distinguishing between sign and non-sign patterns is challenging, as noted by Wang et al. [16], because a single fixed point may only work for some labels. To address this issue, Yang et al. [17] uses a slightly different method in which training for the non-sign models is not required when connecting a CRF model with a single labeled CRF model. The fundamental principle behind handling these frames, whether HMM-based or CRF-based, becomes clear after examining the state-of-the-art methods for dealing with movement epenthesis. Understanding the disparities in speed, hand height, and facial expressions between each phrase is crucial to mastering the technique.

This variability can be quantified by establishing a threshold typically defined by a comparable figure, a minimum, and a maximum. A non-sign frame consists of any movement falling below the minimum or exceeding the maximum point. Despite the similar concept, the main limitation of the current techniques is the considerable time consumption of the Viterbi and related algorithms [18]. Based on the latest analysis, we have noticed that several methods are not suitable for real-time sign detection because they require lengthy computations. Therefore, sign detection algorithms should take into account time complexity.

The study [17] have introduced an approach that decreases the time complexity of the HMM model and is regarded as the foundational model for future sign language recognition.

2.4 Keyframe-based and trajectory-based approaches

Recent studies have explored keyframe extraction and variable-order models for continuous sign language recognition. However, these approaches still face limitations in handling movement epenthesis efficiently, motivating the proposed VOGUE-based framework.

3. PROPOSED MODEL

3.1 Problem formulation

A thorough examination of the movement analysis literature reveals that, given the video sequence, a system tries to identify every gesture that appears in the stream of frames. This issue is presented mathematically in the following manner: Consider the video sequence m, which is a group of frames represented by the symbol $m = \{m_1, m_2, \cdots, m_x\}$, all the methods proposed in the literature try to find the most likely gesture given as a conditional probability $P = (n|m)$. To obtain the value of n, HMM or CRF models are used that try to find the state sequence $n = (n_1, n_2, \cdots, n_y)$. Then the maximum likelihood is obtained using the Eq. (1).

$$\max_{n_1, n_2 \dots} P(n_1, n_2, \dots, n_t | m_1, m_2, \dots, m_t) \tag{1}$$

With this configuration, it is necessary to train the sign parameters to recognize the sign frames and ME frames, often done by using the classic HMM or variants with a fixed or adjustable threshold. Using the CRF, which we mentioned in Section 2, some works perform this training of sign parameters. Following completion of the training, methods such as Viterbi are employed to calculate the maximum likelihood.

The current issue is the computational overhead with this approach, which depends on the number of frames used for likelihood estimates. Finding the ideal indication is more accessible when the complete frame of the video sequence is taken into account, which leads to a low detection rate. When an approach considers the start and finish frames, as is typically the case, ME frames present at either the beginning or the end may cause features to be overlooked. So, an algorithm that can quickly estimate the likelihood is needed for the movement epenthesis problem.

Therefore, the goal of this work is to develop a methodology that is less time-consuming than the state-of-the-art value $O(n^3)$ and that also distinguishes between sign frames and ME frames with a high level of classification accuracy.

3.2 Cluster-based key frame extraction of the sign video sequence

Because signs vary depending on the signer, it might not be easy to recognize them in a video sequence. This dependability causes a video sequence to contain a variable number of frames, which changes the frame rate. When developing a sign recognition system, the system must recognize the signs while not being concerned about frame rate or signer fluctuations. Crisp, purposeful keyframes are required for this. This necessitates a method for separating these distinct frames, which can help with better sign recognition and segmenting of the ME frames. To accomplish movement epenthesis segmentation, key frame extraction is the initial step. The critical frame extraction's overall block diagram is shown in Figure 3.

The video sequence must be processed so that just a few frames are considered before beginning the movement epenthesis segmentation process. This stage is crucial for reducing the number of training samples and enabling the identification of a specific sign in fewer frames. One user's interpretation of a word may differ slightly from another user's interpretation depending on the user making the sign. However, because the computation considers the strokes, it is simple to pinpoint this using the crucial maximum curvature locations. strokes are nothing but breaking a single sign language word into multiple segments, so typically, a stroke is a 2D representation of the hand and expressions.

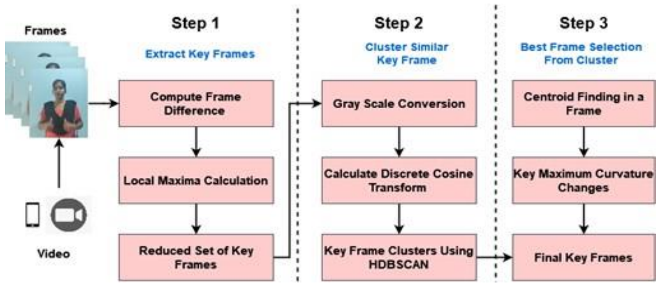
Frames
Video
Step 1
Extract Key Frames
Compute Frame Difference
Local Maxima Calculation
Reduced Set of Key Frames
Step 2
Cluster Similar Key Frame
Gray Scale Conversion
Calculate Discrete Cosine Transform
Key Frame Clusters Using HDBSCAN
Step 3
Best Frame Selection From Cluster
Centroid Finding in a Frame
Key Maximum Curvature Changes
Final Key Frames

Figure 3. Overall process of the key frame extraction

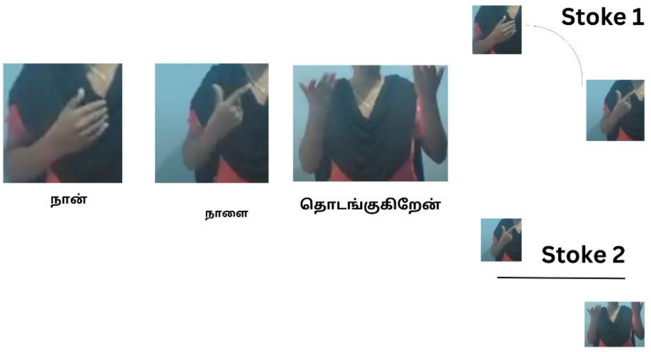

Figure 4. Strokes formed by the sentence "நான் நாளை தொடங்குகிறேன் (I will start tomorrow)"

In theory, this strategy aids in decreasing the training samples since, as seen in Figure 4, there are fewer strokes than words. A stroke is a movement in a video from one frame to the next. If there are 100 frames between the first and second words, it is unnecessary to consider all 100 frames because they all signify the same strokes. The strokes will remain constant regardless of the signer's speed or location on the trajectory, preventing signer-based spatiotemporal fluctuations. Additionally, the sign representation solves the issues raised in the conventional HMM and CRF.

The suggested key frame extraction methodology aids in locating the best frames from the video sequence and the pertinent frames that may be used in additional processing to find the ME frames and signer frames. There are several steps involved in the retroactive frame extraction procedure, and the overall flow of those steps is provided below:

- A window of frames is considered, and the subsequent frame difference value is then calculated between the frames. This quantitative frame difference value will be used as a gauge to identify significant frames. The window's frame with the most significant difference value is chosen to achieve the necessary frames.
- Among the set of frames obtained after the frame difference value undergoes a clustering operation in which the grouping of similar frames happens, the scaling and grayscale conversion of the frame precedes the group. Then the frames undergo a discrete cosine transform that further helps extract informative frames.
- In this work, k-means clustering was used to group visually similar frames. Before clustering, all frames were converted to grayscale and resized to a uniform resolution. Each frame was then transformed using the Discrete Cosine Transform (DCT), and the resulting DCT coefficients were used as feature descriptors. The Euclidean distance metric was employed to measure similarity between frames. The number of clusters was set to k = 6, selected empirically based on preliminary evaluations across multiple sign samples. Frames that did not strongly associate with any cluster centroid were treated as outliers and retained as unique keyframes to ensure that potentially informative frames were not removed during the clustering process.
- Following clustering, the best frames and those that did not fall into any of the clusters form the unique keyframes, which are then strokes segmented again.
- Sign, and ME frames are present following the extraction of unique frames. Therefore, it is crucial to divide the area of interest. In the portrayal of signs, the hand region is essential. Thus, the left- and right-hand areas are considered, and the hand's centroid is used to understand the overall trajectory of hand movement. The changes in hand movement concerning direction and angle are used to segment the global trajectory further. When there is a sudden shift in focus, these changes, known as Key Curvature Maximum Points (KCMP), are indicated. The KCMP points are used to extract additional frames.
- Every continuous sign language representation consists of various signs, generating numerous centroid points that could be uneven. To accomplish the KCMP stroke segmentation, this discontinuity needs to be smoothed out. Approximation must be performed using Bezier or B-spline curves to make the centroid tracking procedure continuous. If the second-order derivative is smooth, the cubic or polynomial B-spline approximation may not work well for dynamic sign language identification because the centroid tracking points might not be equally distributed. So here, nonuniform B-spline approximation is carried out.

3.3 Non-uniform B-spline approximation

A human movement must be captured for sign recognition systems to work. You can accomplish this with hardware, vision, or a hybrid. The hardware-based approach makes use of more expensive, diverse sensors. We have incorporated it into our system because most modern recognition algorithms rely on vision-based information. For poses with varying arm, hand, and face movements, concentration is attained using vision-based data. Along with these benefits, this vision-based strategy is non-intrusive and imposes no limitations on the users. The approximation method that works best to handle the higher-order derivative smoothing as well as the uneven spacing of the points is a non-uniform B-spline. This approximation has the benefit of having several knots that can be used to pull out the curve in any direction without creating a discontinuity. Derivatives can make the non-uniform B-spline more complex, but for the sake of this article, let's stick to the order 4-spline. An open Non-uniform B-spline with control points (here the centroid value) $c_1, c_2, c_3 \dots c_n$ consists of, the knot vector will contain values from $t_0 \dots t_{n+4}$ based on the order, the knot vector values also vary. Assume that there will be four more knots than control points for a 4-spline order.

Based on the control points $P_i - 3, P_i - 2, P_i - 1, P_i$, the B-spline segment $P_i(t)$ is obtained and the expression for obtaining $P_i(t)$ is given by the Eq. (2).

$$P_i(t) = N_i - 3,4(t)P_i - 3 + N_i - 2,4(t)P_i - 2 + N_i - 1,4(t)P_i - 1 + N_i,4(t)P_i \quad (2)$$

where, N denotes the recursive weight functions and $3 \leq i \leq n$ and $t_i \leq t \leq t_{i+1}$. A sample 8-point B-spline curve is shown in the Figure 5.

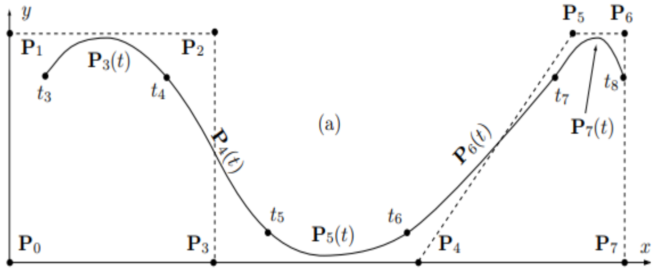

Figure 5. Sample eight-point B-spline curve [19]

3.4 Key Curvature Maximum Points (KCMP)

After approximating the continuous curve, the next step is to find the maximum curvature points, which aid in creating the strokes used for epenthesis frame identification. Mathematically, the KCMP points are derived from the pixels in the trajectory. A pixel is labeled as KCMP in the trajectory if its degree of curvature exceeds the threshold. The neighboring pixels that are considered adjacent to the threshold slope values are computed, as seen in Figure 6. Eq. (3) displays the computation of slope numerically.

$$tan\ tan\ \theta = \frac{ange\ sin\ sin\ y}{ange\ sin\ sin\ x} \quad (3)$$

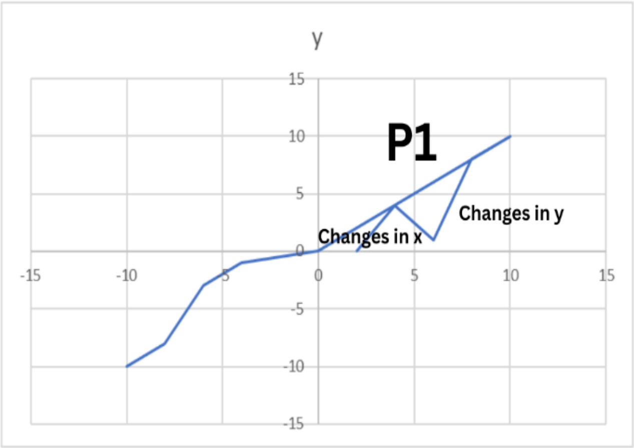

Figure 6. KCMP slope computation

Algorithm 1 illustrates the general process of how the KCMP points are located and provides an algorithmic depiction of KCMP.

The 2D trajectory data from both hands serve as the input for the KCMP extraction, followed by the identification of the region of interest. The area of interest is obtained using the turns of hand movement, which are unaffected by speed and frame rate. Therefore, the speed or frame rate of the signer has no bearing on this process. The main goal is to extract the subunits, and many researchers use various subunits to do this. Pitsikalis et al. [20] developed similar subunit strategies for the German sign language and incorporated the hand as location, shape, and movement as subunits. Similar studies, in which the subunits are described as standard, are carried out by Theodorakis et al. [21] and Aguilera et al. [22]. The segmentation process in our suggested methodology is carried out using the same principle of subunit utilization. The innovative aspect of this study is the way components are used, particularly the slope and direction of the hand, which is an essential aspect in determining the sign language.

Algorithm 1. Key Curvature Maximum Point (KCMP) Selection

```
function KCMP (All points in the Non-Uniform B-spline Approximation)
    for each subsequent point $P_{i-3}, P_{i-}, P_{i-1}, P_i$ do
        Compute $\tan\theta = \frac{\Delta y}{\Delta x}$
        Suppose the sub-segment is $(x_{i-4}, y_{i-4}), (x_i, y_i), (x_{i+4}, y_{i+4})$
        for each sub-segment do
            Compute $\tan\theta_1 = \frac{\partial y}{\partial x}$
            Compute $\tan\theta_2 = \frac{\partial y}{\partial x}$
            Check the condition for marking KCMP:
            if $|\tan\theta_2 - \tan\theta_1| > threshold$
             and $(x_{i+4} - x_i)$ or $(y_i - y_{i-4}$ and $y_{i+4} - y_i)$ then
                $(x_i, y_i)$ is a point in KCMP
            end if
        end for
    end for
end function
```

3.5 Non-manual signs integration

Aside from understanding hand signs and reducing the number of frames, other factors must be considered to identify no-meaning frames in the video sequence, including a factor that heavily depends on the expression and helps convey the word or sentence. Determining the appropriate expression of the sign is therefore crucial. The location of the head, the mouth, and the specific features of the face are used to determine the expressions of the face. The training samples used to convert expressions into emotions and then link those feelings to words are shown in Table 1. The development of a better sign recognition system can benefit immensely from this.

We must recognize the face and pay attention to features such as the eyebrows, mouth, eyes, and lips to separate emotions from expression. Since this does not help identify ME frames, we did not execute the procedure to encode emotions into sentences in this work. To determine whether the frames are an aid in the recognition of signs, it is crucial to consider both the expression and the information of the hand. The HOG classifier uses the histogram of directed gradients to extract key points and identify the region of interest. For testing reasons, these expressions are flags within the Tamil sentence structure, indicating a 0 for no emotion and a 1 for the sentiment. If a feeling is detected, this frame may belong to the sign frames and should be removed from the ME frames.

Table 1. Facial expression and their associated emotions in Tamil

Feature Expression	Emotion Associated	Expression in English	Possible Sentence Form
Raised Eyebrows	Question expecting answers in the form of Yes/No	Do you want this?	இது வேணுமா?
Lowered Eyebrows	Question expecting answers to know the place, person etc.	Why not play?	ஏன் விளையாட்டு இல்லை?
Wide Lips	Happiness	I have received a promotion.	எனக்கு பதவி உயர்வு கிடைத்துள்ளது
No Head Movement	Negative	Do not go there.	அங்கே போகாதே
Mouth opens to an oval	Surprise	Oh! / Wow!	ஆஹா

The algorithm 2 shows the overall procedure involved in identifying and decoding the facial expression of an emotion.

Algorithm 2. Face Expression Identification

```
function FaceExpression
    Boundary1 = HOG-SVM classifier of eye()
    Boundary2 = HOG-SVM classifier of mouth()
    Boundary3 = HOG-SVM classifier of head()
    KeyPoint_Eyebrow = Extract_Eyebrow(Boundary1)
    KeyPoint_Mouth  = Extract_Mouth(Boundary2)
    KeyPoint_Head   = Extract_Head(Boundary3)
    if KeyPoint_Eyebrow moves upward then
        Set flag = 1
    else if KeyPoint_Head moves back and forth then
        Set flag = 1
    else if KeyPoint_Eyebrow moves downward then
        Set flag = 1
    else if KeyPoint_Lips == wide then
        Set flag = 1
    else if KeyPoint_Mouth opens to oval then
        Set flag = 1
    else
        Set flag = 0
    end if
end function
```

3.6 VOGUE (Variable order and gapped HMM for unstructured elements) model-based movement epenthesis segmentation

The VOGUE model [19, 23] is employed in this work for the segmentation of movement epenthesis in continuous sign language. VOGUE is particularly advantageous because it enables the detection of ME frames in linear time, significantly reducing computational overhead compared to conventional HMM- and CRF-based segmentation approaches.

Following the key frame extraction procedure described in Section 3.2, a set of distinct frames is obtained and represented as:

$$u = u_{1,u_2,u_3} \cdots \cdot u_n$$

This set contains both meaningful sign frames and potential ME frames. The objective is to accurately separate ME frames from this sequence.

Prior studies [24, 25] eliminate filler gestures using exhaustive search-based procedures, which are computationally intensive. In contrast, the proposed approach integrates all relevant keyframes into a unified sequence, enabling VOGUE to isolate ME frames efficiently using probabilistic modeling.

Traditional Markov models assume that the next state in a sequence depends on a fixed number of preceding states. However, in continuous sign language, the length of contextual dependency varies depending on signer speed, trajectory, and stylistic differences. VOGUE addresses this limitation by employing a variable-order Markov model, which automatically adapts the context length based on the observed data. This ability to adjust context enables the model to capture subtle sign transitions more accurately.

VOGUE incorporates gap modeling, allowing it to represent variable-length transitions between strokes. Since ME frames commonly appear as irregular gaps between meaningful gestures, gap modeling plays a crucial role in identifying epenthesis boundaries.

The process of movement epenthesis segmentation in VOGUE involves the following steps:

- The Variable Gap Sequencer (VGS) extracts frequent subsequences from the keyframe sequence.
- A variable-order Markov model is constructed using these variable-length subsequences.
- During segmentation, the likelihood of each frame is computed in the form of a log-ratio. The model order is increased iteratively to identify ME frames based on variations in the probability distribution.

3.6.1 Advantages of VOGUE over HMM and CRF models

The VOGUE model offers several advantages compared to conventional HMM and CRF methods for movement epenthesis segmentation. Classical HMMs rely on fixed-order dependencies, making them inadequate for sequences in which contextual requirements vary dynamically, as is typical in continuous sign language. Although variable-duration HMMs account for duration variability, they still require explicit duration modeling, which becomes unreliable when ME frames durations are inconsistent and unstructured.

In contrast, VOGUE automatically captures variable-length contextual dependencies through the use of a context tree. It also learns gap-length distributions directly from the mined frequent sequences, enabling the model to distinguish between meaningful sign strokes and non-sign transition movements without the need to explicitly model ME frames.

Furthermore, HMM and CRF approaches typically rely on Viterbi decoding, which has a cubic time complexity with respect to sequence length. This presents a major computational bottleneck for real-time systems. VOGUE performs likelihood estimation in linear time, making it significantly more efficient and suitable for real-time continuous sign language recognition applications [26].

The capability to model unstructured segments, handle variable-duration gaps, and compute probabilities efficiently constitutes the primary motivation for adopting VOGUE as the segmentation framework in this work.

Algorithm 3. Variable Key Frame Sequence Mining (VKFSM)

```
function VKFSM
    Input:
        Maximum Gap Allowed (MG)
        Maximum Sequence Length (L)
        Minimum Frequency Threshold (MT)
    for every element in L do
        Find the sequences with frequencies of length 1
        for all elements with length 1 do
            Extend sequence length to 2
            Obtain gap length distribution
            for each frequent sequence do
                Record symbol distribution
            end for
        end for
    end for
end function
```

After obtaining the mined sequence set using the VGS algorithm, these are used for the building of the VOGUE model. Every non gap frame in the keyframe is represented as a state here, and the gap length and symbol distribution are considered when adding any state in between two states. This is particularly useful in extracting the ME frames because the gaps generally denote the end of the first word, which could be a part of the ME frames, and also because the symbol distribution helps us conclude regarding the not-so-ME frames.

The VOGUE model must learn two things to perform ME segmentation. First, the set of all Tamil sentences captured as sign videos is named as data. The procedure for learning these training data is based on the probability distribution algorithm. The context tree is then used to learn the strokes sequences as well as the mined sequence of the sign language. Given past data, the model assigns future key frame probabilities based on the given past data. The calculation performed internally and the probability calculation for the next frame are given by Eq. (4). The conditional distributions of the VOGUE model take the form $P(sign\ of\ keyframe \mid past\ history)$. Here, the approach we are following is to group all the sign frames of a word so that when the end of the word is reached before the start of the next word, all these signs are not recognized, so these will be marked as ME frames. So, the likelihood computation is done using Eq. (4).

The procedure for learning the VOGUE model is shown in algorithm 4. Since this method aims to aggregate the frames that make up signs, it is not necessary to specific model ME frames. This methodology is effective since the most challenging part of segmenting ME frames is modeling them.

$$P(sign\ of\ keyframe \mid past\ history) = \frac{1}{2} + \frac{Number\ of\ stroke\ occurrences\ in\ training\ data\ after\ context}{Total\ occurrences\ of\ context\ in\ training\ data} \tag{4}$$

Algorithm 4. VOGUE-Based Momentum Epenthesis Segmentation

```
function VOGUE
    Input:
        Test Sentence: Length N, set of keyframes
        S = {s1, s2, …, sn}
    Output:
        Sign Symbols [] = sign words list[]
    Parameters:
        State_Prev    = 0
        State_Current = starting index
        State_Next    = 2
    function ContextTree(<word, vocabulary>)
        for each word in the word vocabulary do
            Update state probability and compute
            $P(every\ sign\ of\ keyframe \mid past\ history\ of\ each\ sign$
        end for
    end function
    for i = 1 to N do
        Current State Keyframe = Current State Keyframe + State_Current
        Previous State Keyframe = Previous State Keyframe + State_Prev
        Compute Log Ratio
        if Log Ratio > Threshold then
            Set Sign Flag = 1
            Add the current frame to the sign frame group
        end if
    end for
    return Sign Frames []
end function
```

4. RESULT AND DISCUSSION

4.1 Data set

Ten videos created and used based on the Tamil sign language are used to test the separation between ME and sign frames. The dataset used in this study consists of 30 video samples derived from 10 distinct Tamil sign language sentences, each performed by three different signers. A train–test split of 70% and 30% was employed, respectively, while ensuring signer independence by assigning different signers to the training and testing sets. This cross-signer evaluation strategy was adopted to assess the generalization capability of the proposed method across variations in signer style, speed, and trajectory. Since publicly available datasets for Tamil Sign Language are limited, the dataset was custom-recorded for this study, and its size reflects the practical constraints of data collection in low-resource sign languages. Three separate samples are produced using these ten sentences, each signed by a different signer whose pace, trajectory, and hand shape differ. In terms of sign exposure, only a small number of phrases exactly matched the three signers, and in a few instances also had wider stylistic variances.

4.2 Cluster based on frame extraction results and discussion

The reduction of training time is the main objective of cluster-based frame extraction. Figure 7 illustrates how few important frames were retrieved during KCMP frame extraction. The frame ratio is calculated as a ratio between the total number of keyframes produced after the cluster extraction and the total number of frames to comprehend the number of frames reduced quantitatively. The sequence in Figure 7 was given a frame ratio of 0.0159. Because the signer becomes more aware of each word and casually keeps the remaining spaces, the turning points of the hand trajectory viewed as the final step for collecting the keyframes made sense and made our job more manageable. Figures 8 and 9 represent the point extracted KCMP on the trajectory and the equivalent frames extracted.

The suggested method for decreased key frame extraction yielded good results with fewer frames. However, it is crucial to understand whether or not this decrease in frames affects how well indications can be detected. So, sign identification is validated using some of the matching algorithms like scale-invariant feature transform (SIFT), speed-up robust feature (SURF), robust independent elementary features (BRIEF), oriented FAST, rotated BRIEF (ORB), and the results obtained are tabulated in Table 2.

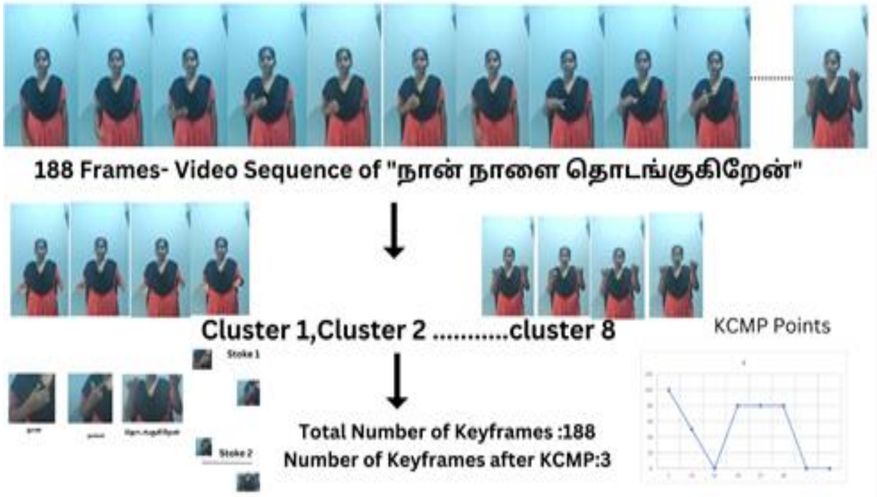
188 Frames- Video Sequence of "நான் நாளை தொடங்குகிறேன்"
Cluster 1,Cluster 2 ....cluster 8
KCMP Points
Stoke 1
Stoke 2
Total Number of Keyframes :188
Number of Keyframes after KCMP:3

Figure 7. Cluster-based frame extraction of a hand trajectory for sign identification

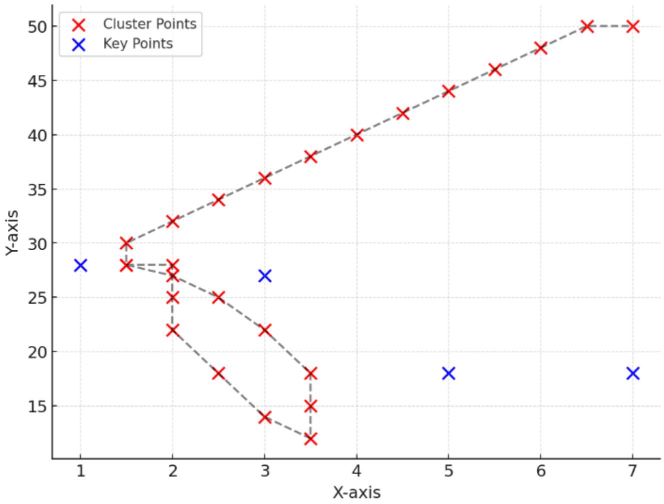

Figure 8. Cluster based trajectory for the word “நான் நாளை தொடங்குகிறேன் (I will start tomorrow)”

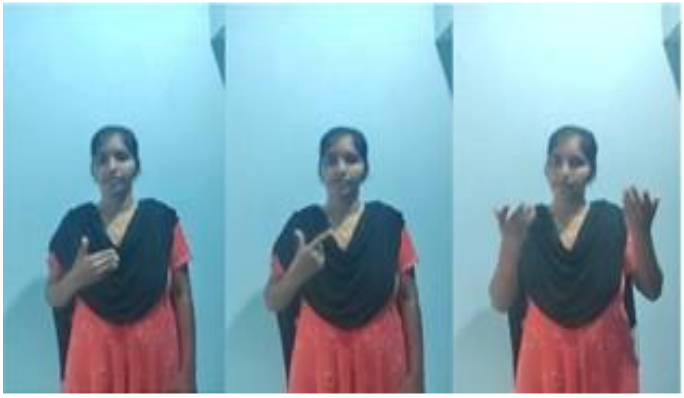

Figure 9. Key frame extraction for the sign “நான் நாளை தொடங்குகிறேன் (I will start tomorrow)”

Table 2. Tabulated values showing the state-of-the-art matching algorithms for sign recognition without cluster-based key frame extraction and with cluster-based key frame extraction

	Without Cluster-Based Key Frames				**With Cluster-Based Key Frames**			
Matching Algorithms	**Per Sign Frames**	**Key Frames for Sign**	**Frame Ratio**	**Accuracy**	**Per Sign Frames**	**Key Frames for Sign**	**Frame Ratio**	**Accuracy**
SIFT	18	18	1	68.92%	18	1	0.05	58.92%
SURF	18	18	1	73.72%	18	1	0.05	83.76%
ORB	18	18	1	75.76%	18	1	0.05	82.17%

Table 3. State-of-the-art results in terms of error rate for various methods

Model	**Sign Error Rate**
Dynamic Time Warping	90.83%
Hidden Markov Model	82.70%
Conditional Random Field (Fixed Threshold)	66.04%
Conditional Random Field (Short Sign Detector)	67.08%
Conditional Random Field (Non-Sign Patterns Labelling)	59.79%
VOM Model [25]	6.8%

It is observed that the accuracy of certain feature-based matching methods, such as SIFT, decreases after applying cluster-based key frame extraction. This reduction occurs because SIFT relies heavily on dense local keypoints, and the removal of intermediate frames results in fewer distinctive feature points available for matching. Consequently, the descriptor becomes less discriminative when only one or two frames represent a complete strokes. In contrast, SURF and ORB remain more stable or even improve in accuracy because they use more robust gradient-based and binary descriptors that tolerate reduced frame density. These results highlight a trade-off between frame reduction and the sensitivity of different feature extractors, explaining why SIFT shows decreased performance in the clustered condition.

Next, the performance of the VOGUE model is verified using testing sentences that were not part of the training phase. The key frame sequence then goes through the probability computation for every iteration in the context tree before a particular frame is considered for grouping as sign frames. Finally, the state-of-the-art comparisons in terms of the sign error rate are verified with our methodology, and the tabulation for the same is depicted in Table 3. Because this data set was created specifically for this study and is much smaller in size, the results achieved with our methodology cannot be compared to those obtained with other state-of-the-art methods.

Thus, these data are displayed to perform a comparative analysis, but they are not displayed to highlight the superiority of our model over competing models. The mistake rate discovered using our suggested strategy is 5.3% with an overall accuracy of 86.78%. Because most of the learning occurs via the Viterbi algorithm, another significant issue with ME frames segmentation is that the temporal complexity of the suggested approaches is cubic as $O(n^3)$. However, because the estimation is based on probability and logarithmic ratio, this VOGUE model executes the learning in linear time. As a result, $O(n)$ is the time complexity.

5. CONCLUSIONS AND FUTURE WORK

The focus of this project was to develop a new methodology for solving the movement “epenthesis” problem in the context of sign language recognition, to create a more effective sign language recognition system. The proposed approach involves extracting relevant frames that contain information about hand shape, movement, and facial expression. The methodology utilizes a clustering technique to group similar frames, followed by the Key Frame Capturing (KCMP) process to identify key frames. A VOGUE-based learning model is then used to learn sequences and compute probabilities to determine the relevance of the current frame to the previous frame. If there is a significant variation in the log ratio value, it is understood that the frame does not belong to the current word; it may be a movement epenthesis frame or a sign for the next word. The proposed model demonstrated improved performance, with a sign error rate of around 5%. The error rate remained consistent despite changes in speed and direction within the sequence. The experimental results indicate that the proposed method achieves competitive accuracy and improved computational efficiency compared to established baseline techniques, within the constraints of the evaluated Tamil Sign Language dataset. These findings demonstrate that the VOGUE-based framework is effective for real-time movement epenthesis segmentation without overstating generalization beyond the current dataset. One potential enhancement could involve the use of a Bayesian deep learning network for matching and recognition.

ACKNOWLEDGMENT

The authors express their heartfelt thanks to Dr. R. I. Minu, Professor at SRM Institute of Science and Technology, Chengalpattu, India, for their constant support, encouragement, and help rendered to complete this research work.

REFERENCES

[1] Theodorakis, S., Pitsikalis, V., Maragos, P. (2010). Model-level data-driven sub-units for signs in videos of continuous sign language. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, pp. 2262-2265. https://doi.org/10.1109/ICASSP.2010.5495875

[2] Myers, C., Rabiner, L. (1981). A level building dynamic time warping algorithm for connected word recognition. In IEEE Transactions on Acoustics, Speech, and Signal

Processing, 29(2): 284-297. https://doi.org/10.1109/TASSP.1981.1163527

[3] Lichtenauer, J.F., Hendriks, E.A., Reinders, M.J. (2008). Sign language recognition by combining statistical DTW and independent classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11): 2040-2046. http://doi.org/10.1109/TPAMI.2008.123

[4] Li, W., Luo, Z., Xi, X. (2020). Movement trajectory recognition of sign language based on optimized dynamic time warping. Electronics, 9(9): 1400. http://doi.org/10.3390/electronics9091400

[5] Fang, G., Gao, W., Zhao, D. (2006). Large-vocabulary continuous sign language recognition based on transition-movement models. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(1): 1-9. http://doi.org/10.1109/TSMCA.2006.886347

[6] Yang, R., Sarkar, S., Loeding, B. (2009). Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3): 462-477. http://doi.org/10.1109/TPAMI.2009.26

[7] Manuel, M., Menon, A.S., Kallivayalil, A., Isaac, S., KS, D.L. (2021). Automated generation of meeting minutes using deep learning techniques. International Journal of Computing and Digital System, 12(1): 109-120. http://doi.org/10.12785/ijcds/1201010

[8] Vogler, C., Metaxas, D. (1997). Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. In 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, Orlando, FL, USA, pp. 156-161, IEEE. http://doi.org/10.1109/ICSMC.1997.625741

[9] Lee, H.K., Kim, J.H. (1999). An HMM-based threshold model approach for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(10): 961-973. https://doi.org/10.1109/34.799904

[10] Kelly, D., McDonald, J., Markham, C. (2009). Recognizing spatiotemporal gestures and movement epenthesis in sign language. In 2009 13th International Machine Vision and Image Processing Conference, Dublin, Ireland, pp. 145-150. http://doi.org/10.1109/IMVIP.2009.33

[11] Kim, J.B., Park, K.H., Bang, W.C., Bien, Z.Z. (2002). Continuous Korean sign language recognition using gesture segmentation and hidden Markov model. In 2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No.02CH37291), Honolulu, HI, USA. http://doi.org/10.1109/FUZZ.2002.1006741

[12] Yang, W., Tao, J., Ye, Z. (2016). Continuous sign language recognition using level building based on fast hidden Markov model. Pattern Recognition Letters, 78: 28-35. http://doi.org/10.1016/j.patrec.2016.03.030

[13] Choudhury, A., Kumar Talukdar, A., Kamal Bhuyan, M., Kumar Sarma, K. (2017). Movement epenthesis detection for continuous sign language recognition. Journal of Intelligent Systems, 26(3): 471-481. http://doi.org/10.1515/jisys-2016-0009

[14] Lafferty, J., McCallum, A., Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pp. 282-289.

[15] Yang, H.D., Lee, S.W. (2010). Simultaneous spotting of signs and fingerspellings based on hierarchical conditional random fields and boostmap embeddings. Pattern Recognition, 43(8): 2858-2870. https://doi.org/10.1016/j.patcog.2010.03.007

[16] Wang, S.B., Quattoni, A., Morency, L.P., Demirdjian, D., Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, pp. 1521-1527. http://doi.org/10.1109/CVPR.2006.132

[17] Yang, H.D., Sclaroff, S., Lee, S.W. (2008). Sign language spotting with a threshold model based on conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(7): 1264-1277. http://doi.org/10.1109/TPAMI.2008.172

[18] Schmidt, C., Koller, O., Ney, H., Hoyoux, T., Piater, J. (2013). Using viseme recognition to improve a sign language translation system. In Proceedings of the 10th International Workshop on Spoken Language Translation: Papers.

[19] Koller, O., Bowden, R., Ney, H. (2016). Automatic alignment of hamnosys subunits for continuous sign language recognition. LREC 2016 Proceedings, pp. 121-128.

[20] Pitsikalis, V., Theodorakis, S., Vogler, C., Maragos, P. (2011). Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In CVPR 2011 Workshops, Colorado Springs, CO, USA, pp. 1-6. http://doi.org/10.1109/CVPRW.2011.5981681

[21] Theodorakis, S., Pitsikalis, V., Maragos, P. (2014). Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image and Vision Computing, 32(8): 533-549. http://doi.org/10.1016/j.imavis.2014.04.012

[22] Aguilera, A.M., Aguilera-Morillo, M.C. (2013). Comparative study of different B-spline approaches for functional data. Mathematical and Computer Modelling, 58(7-8): 1568-1579. http://doi.org/10.1016/j.mcm.2013.04.007

[23] Zaki, M.J., Carothers, C.D., Szymanski, B.K. (2010). Vogue: A variable order hidden Markov model with duration based on frequent sequence mining. ACM Transactions on Knowledge Discovery from Data (TKDD), 4(1): 1-31. http://doi.org/10.1145/1644873.1644878

[24] Wilcox, L.D., Bush, M.A. (1992). Training and search algorithms for an interactive wordspotting system. In [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, USA, pp. 97-100. https://doi.org/10.1109/ICASSP.1992.226111

[25] Geetha, M., Kaimal, M.R. (2018). A 3D stroke based representation of sign language signs using key maximum curvature points and 3D chain codes. Multimedia Tools and Applications, 77(6): 7097-7130. https://doi.org/10.1007/s11042-017-4624-y

[26] Duraimutharasan, N.K.B., Sangeetha, K. (2023). Machine learning and vision based techniques for detecting and recognizing Indian sign language. Revue

d'Intelligence Artificielle, 37(5): 1361-1366. https://doi.org/10.18280/ria.370529