VNU Journal of Science: Comp. Science & Com. Eng., Vol. 41, No. 1 (2025) 54–68

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Sign Language Representation using Virtual Characters with 3D Animation

Thi Duyen Ngo*, Duc Hoang Long Nguyen, Hai Long Luong

VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Received 07 October 2024
Revised 15 January 2025; Accepted 05 March 2025

Abstract: Sign language is a communication system that encompasses bodily gestures, primarily utilized within the deaf community. Due to its limited prevalence, information from books, newspapers, and videos is often not translated or represented in sign language. This situation creates challenges for deaf individuals in acquiring information, as well as in their learning and interactions with hearing individuals. Historically, the conversion between spoken language and sign language relied entirely on interpreters, a limited resource that is not always readily available. Currently, employing technology to convert spoken language into sign language presents a modern and convenient alternative. This linguistic conversion typically involves two steps: first, converting spoken language into text that adheres to the grammatical structure of sign language; second, representing this text through the corresponding gestures. This paper proposes a method for representing sign language using 3D characters to address the latter step. The method constructs a 3D skeleton motion for each word or phrase from input text in sign language grammar. Subsequently, the motion data of words is processed and interconnected to animate a 3D virtual character for the complete sentence representation. We have applied the proposed method to represent Vietnamese Sign Language (VSL) using 3D virtual characters. The results were assessed by experts in sign language, yielding promising findings that suggest the practical applicability of the proposed methodology.

Keywords: Vietnamese Sign Language, Sign language representation, 3D animation, Virtual character.

1. Introduction

According to the World Health Organization (WHO) [1], it is projected that by 2024, over 5% of the global population, approximately 430 million individuals, will require rehabilitation to manage the impacts of hearing loss.

*Corresponding author.
E-mail address: duyennt@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.3768

In Vietnam, this figure is estimated to be around 2.5 million people. Deaf individuals often lack proficiency in spoken language and do not primarily use it as their means of communication; instead, they rely on sign language, utilizing bodily gestures to convey information. However, sign language is not widely used outside the deaf community, and resources such as books, newspapers, and videos frequently do not accommodate sign language. Sign language possesses a entirely different grammar from spoken language, complicating the reading of written materials for deaf individuals. In educational settings, deaf students face significant challenges in communication and learning due to the limited number of teachers proficient in sign language. These factors underscore the need for a technology-based solution for translating text into sign language.

Sign languages exhibit considerable variation across different countries and even within regions of the same country. This diversity presents a challenge, as it complicates the broad application of sign language research. However, it also serves as a catalyst, prompting countries to undertake research tailored to the specific needs of their native deaf communities. There are some research developed aimed at translating from spoken language to sign language in several languages such as American [2, 3], India [4, 5], Pakistan [6], Brazil [7], Sinhala [8] and Arabic [9–11]. In Vietnam, while there are some studies focused on the field of sign language, these have primarily addressed the issue of VSL recognition [12, 13], which involves translating sign language into spoken language. Conversely, to our knowledge, only one study has proposed a solution for translating from spoken language to sign language [14].

Methods for translating spoken language into sign language for the deaf community generally proceed in two fundamental stages. Initially, spoken language input, whether in text or audio form, is converted into text with sign language grammar. Subsequently, the processed text is represented in various formats, such as virtual characters [15], motion graphs [16], or 2D images and videos [17]. Among these methods, virtual characters provide several notable advantages; they offer clearer and more expressive sign language representations and exhibit higher storage efficiency compared to videos [18]. The motion synthesis technique that most authentically conveys the realism of virtual characters is data-driven [19]. Previous research adopting this approach has typically yielded natural and realistic motion [20–24]. However, these studies have frequently encountered challenges related to the cost of dataset construction, which is predominantly driven by the substantial expenses associated with the motion capture process. Motion capture typically requires specialized equipment, such as markers [22], multi-camera systems [23, 24] or wearable devices like Cybergloves [20, 21].

In this study, we propose a method for representing sign language through the use of 3D virtual characters, with a focus on reducing the costs associated with motion capture while maintaining realistic and natural movements. The method can be applied to any language with input being a text that has been processed in the sign language grammar and result in a 3D representation of the sign language. The approach begins by constructing a 3D skeleton motion for each word in the text. This motion data is subsequently interconnected and utilized to animate a virtual character, thereby creating a 3D representation of sign language. This method has been successfully applied to VSL, resulting in the development of a skeleton motion dataset and animations for VSL using a 3D virtual character. This advancement not only enhances the visual representation of sign language but also broadens the potential for inclusive communication within the deaf community.

The rest of the paper is structured as follows: Section 2 surveys the research relevant to our

study. In Section 3, we present a detailed explanation of the proposed method. Section 4 outlines the findings from our evaluation of the video representing sign language generated by the proposed method. The concluding Section 5 recaps the paper's content and the discussed method.

2. Related work

In the task of representing sign language from text input in sign language grammar, three common methods are employed: the use of virtual characters, the application of motion graphs, and the utilization of 2D images and videos. Each method possesses its own advantages and disadvantages. However, recent studies have increasingly favored the use of virtual characters as the optimal approach due to their ability to authentically and accurately depict movements, which is a critical factor in the representation of sign language.

Yosra Bouzid and Mohamed Jemni [2] presented a method for generating 3D animation sequences from SignWrititng notation for American sign language (ASL). The input is the XML format of SignWrititng which is called SignWrititng Markup Language. After processing, the notation is translated to Sign Modeling, which is then automatically interpreted by the WebSign player [25].

In 2015, Diego et al. [7] developed a synthesis system that interprets XML inputs describing hand gestures and converts them into a vector of configuration parameters. These parameters are then used to animate a 3D avatar, representing the Brazilian Sign Language. A paper published the same year by Punchimudiyanse et al. [8] proposed a multi-facet 3D avatar and an animation system for Sinhala Sign Language(SSL) that allows sign movements to be defined and animated without the need for motion capture hardware or video sequencing. Testing with a vocabulary of 200 signs and 40 finger-spelling signs showed that the system effectively animated various sentence types in SSL, demonstrating its flexibility and potential for broader applications.

Kaur et al. [4] proposed an automation system that generates HamNoSys for Indian Sign Language (ISL) words. This is accomplished by converting the Hamburg Notation System (HamNoSys) [26] into Signing Gesture Mark-up Language (SiGML)[27], which then is processed to animate the corresponding signs. The system contains a database of approximately 210 HamNoSys symbols. Also using HamNoSys and SiGML, Bhavinkumar et al. [5] proposed ES2ISL, a system that converts English speech to ISL.

A method for Arabic Sign Language (ArSL) was proposed by Al-Barahamtoshya et al [9]. It records words using a voice module and then converts them into ArSL using a transition module. The suggested method converts the text into ArSL by using an Arabic language model and a set of transformational rules. However, the paper did not mention the sign presentation and the method of animating the ArSL sign.

Muhammad Sanaullah et al. [6] proposed a real-time automatic translation system called Sign4PSL that converts English text into Pakistan Sign Language (PSL) using a virtual signing character. The proposed method includes converting words to HamNoSys Notation transcription according to sign specifications. The transcription then is converted to SiGML tags, which are sent to the AnimGen client-server at UEA for the Avatar signing commands extraction.

In Vietnam, the only attempt at representing VSL is by. Luyl Da Quach et al [14]. They addressed the challenge of converting Vietnamese television news into 3D sign language. The proposed method uses ID3 to transform input sentences into VSL grammar sentences. The translated text is then processed to generate a SiGML file which is used for JA Signing [28] to

create HamNoSys codes to animate a 3D avatar.

The utilization of notation and markup languages, such as SignWriting [2], XML [7], HamNoSys and SiGML [4–6, 14] is a common method for depicting gestures in sign language. These structured input formats facilitate the systematic translation of text into corresponding animated sign language visuals, thereby enabling automation in the animation process. However, this approach presents significant drawbacks; for instance, reliance on specific markup languages can limit the scalability of the vocabulary due to the necessity of involving sign language experts. Furthermore, there have been studies employing a data-driven approach that eliminates the reliance on notation and markup languages. However, these methods still entail high costs during the motion capture process and require specialized equipment, such as systems involving multiple cameras [24] or Cybergloves [20, 21]. The method proposed in this paper introduces a pipeline to represent sign language using virtual characters, addressing these limitations by minimizing the costs associated with the dataset construction.

3. Proposed Method

To address the challenge of representing sign language from text in sign language grammar, our proposed method involves constructing 3D skeleton motion for each word or phrase in the input sentence. Subsequently, this motion data is processed and interconnected to create an animation that represents a complete sentence. The generated data for the words or phrases will be stored in a dataset for future use. This approach is advantageous due to its flexibility in assembling words into coherent sentences. Consequently, creating a complete sentence from a dataset of motion data for individual words and phrases is significantly more feasible than developing a motion dataset for complete sentences. Additionally, expanding the dataset

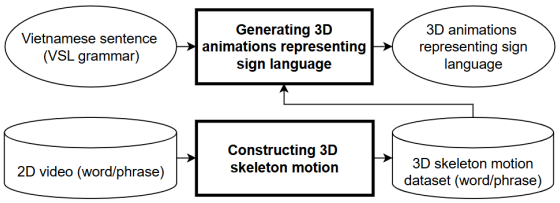

Figure 1. Overview pipeline for representing sign language with 3D virtual characters.

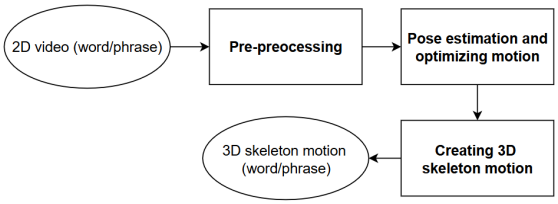

Figure 2. Overview pipeline for constructing 3D skeleton motion for words.

can be accomplished easily by using video inputs that demonstrate sign language. Therefore, we propose a method consisting of two steps, as illustrated in Fig. 1: constructing 3D skeleton motion for words and generating 3D animations representing sign language.

3.1. Constructing 3D Skeleton Motion for Words

To construct 3D skeleton motions for a word from a 2D video, we employ a methodology outlined in our previous work [29] that consists of three steps, as illustrated in Fig. 2. The input for this process is a 2D video of a person performing a word or phrase in sign language. The output of this process is 3D skeleton motion, which can subsequently be utilized to animate a 3D character representing sign language. This approach facilitates the expansion of the 3D skeleton motion dataset by enabling the collection or creation of additional 2D videos of sign language. However, a challenge associated with 2D videos is the variability in properties such as sources, creators, frame rates, and resolution. Consequently, the proposed method also addresses this challenge. The effectiveness

of the proposed 3D motion construction approach was evaluated in our previous study [29], where 20 evaluators assessed the similarity between VSL signs performed by virtual characters and those performed by real people. An average score of 3.93 on a Likert scale of 1 to 5 demonstrates the robustness and suitability of the proposed 3D motion construction approach for sign language representation.

3.1.1. Pre-processing

The use of diverse 2D video sources depicting sign language can lead to inconsistencies in frame rates, as these videos are often produced by different individuals at various times or recorded using different devices. This variability can result in animations where certain signs are displayed more quickly or slowly than intended. To address this issue, we preprocess the 2D sign language videos to achieve a uniform frame rate of 30 fps. This standardization facilitates the synchronization of sign execution speeds and allows for precise control over the movements of the virtual character. Specifically, each new video V' is generated from the original video V using the following equation:

$$V'_i = V_{[\frac{i \times fps}{fps'}]} \qquad \forall i \in [0, \frac{|V| \times fps'}{fps}] \quad (1)$$

where V_i and V'_i are the i-th frames of the videos V and V', respectively, and the frame rates of videos V and V' are fps and fps′, which is set to 30 in this research.

3.1.2. Pose Estimation and Optimizing Motion

After adjusting the frame rate of the videos to the same level, key points on the body are identified for each frame and aggregated into complete postures, a process called pose estimation. In this study, we employed OpenPose [30] to identify 2D key points across the entire body (Fig. 3) and specifically on the hands (Fig. 4). Following experiments with different configurations, we have adopted the

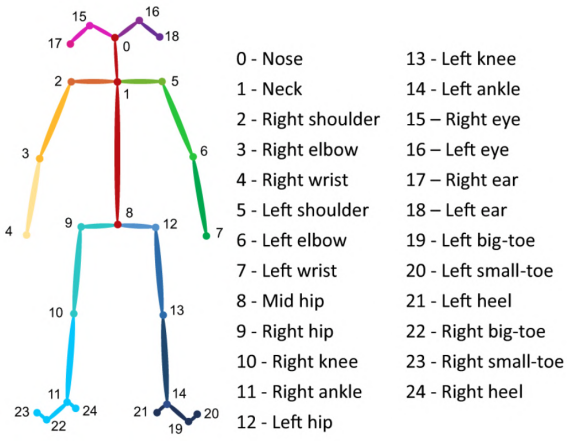

Figure 3. The key points on the body are identified using OpenPose [30].

BODY_25 format, which identifies 25 key points across the whole body, as depicted in Fig. 3, to extract 2D motion from sign language videos. Since our research focuses on representing sign language, we only collect the motion of key points on the body from the waist up.

To optimize 2D motion data, redundant movements are pruned from the 2D motion data, which improves the performance of the 3D sign language dataset construction and the accuracy of the final output. The actions of raising and lowering the hand, which frequently occur in sign language videos but do not convey meaning, are identified and removed. This is achieved by detecting the time intervals where actual signs are being performed, based on the solution proposed by Amit Moryossef et al. [31], which utilizes machine learning techniques with over 90% reported accuracy. By eliminating these non-essential movements, the computational resources required for subsequent processing steps are significantly reduced.

First, the variation in motion between consecutive frames is computed using equation 2.

$$F(P)_t = \|P_t - P_{t-1}\|_2 * fps \quad (2)$$

where P is the 2D motion data, t is the frame index, and fps is the frame rate of the input video.

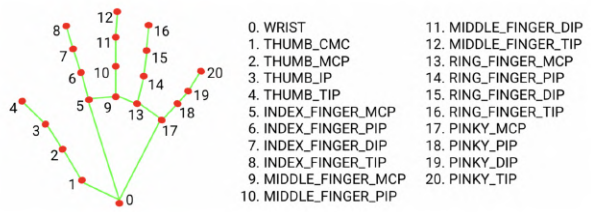

Figure 4. The key points on the hand are identified using OpenPose [30].

The points P that cannot have their coordinates determined at frame t have values $F(P)_t$ and $F(P)_{t+1}$ defaulted to 0.

The pre-trained sign language detection model [31] takes as input the variation of the 2D motion data of the entire body (excluding the hands) from consecutive frames and outputs the probabilities of the presence and absence of sign language gestures between the frames. The determination of the presence of sign language is given by equation 3 where P is the motion data and M is the sign language detection model.

$$signing(P) = argmax(M(F(P))) \tag{3}$$

The $signing(P)$ receives a value of 1 or 0, corresponding to the "presence" and "absence" of sign language gestures respectively. To ensure the continuity of motion, frames from the first frame to the last frame that receives a result of 1 are considered to contain sign language gestures, and the 2D motion data of these frames will be processed in the subsequent steps. In our analysis, we found that the model designed for recognizing hand-raising motions demonstrated a high level of efficacy. However, its performance in detecting hand-lowering motions was notably less effective. To address this limitation, we revised the original equation 3 to the updated equation 4:

$$signing(P) = argmax(M(F(P))) \;\; OR \;\; argmax(reverse(M(reverse(F(P))))) \tag{4}$$

The function *reverse* is defined as the operation that reverses the order of elements within an array. By reversing the video sequence, the motion of lowering the hand becomes analogous to that of raising the hand. This modification significantly enhances the efficacy of sign language recognition, thereby optimizing the detection of hand movements

3.1.3. Creating 3D Skeleton Motion

The concluding stage of the process necessitates the coordinates of the hip points (specifically points 8, 9, and 12 as depicted in Fig. 3) as essential input parameters. However, a notable limitation arises from the fact that videos depicting sign language typically only capture the upper body of the performer. This results in a significant number of videos where the identification of these critical hip points is rendered infeasible. To effectively mitigate this challenge, we have devised a methodology [29] to calculate the coordinates of the three hip points based on the coordinates of the three shoulder points (points 1, 2, and 5 as illustrated in Fig. 3). Their coordinates were estimated using the following equations:

$$H_m = S_m + D_m \times L, \tag{5}$$

$$H_l = H_m + D_l \times L, \tag{6}$$

$$H_r = H_m + D_r \times L. \tag{7}$$

The meanings of the symbols used in this section are provided in Table 1. In these equations, L, which represents the shoulder size, is computed as follows:

$$L = \|S_l - S_m\|_2 + \|S_r - S_m\|_2. \tag{8}$$

To determine the appropriate distance ratios for these estimations, an empirical analysis was conducted on a diverse set of videos. Based on the results, a table of fixed distance ratios between key points relative to shoulder width was established (Table 2). These ratios

Table 1. Definitions of the symbols used in pose estimation

Symbol	Definition
L	Shoulder size
S_m	Middle shoulder coordinate
S_l	Left shoulder coordinate
S_r	Right shoulder coordinate
H_m	Middle hip coordinate
H_l	Left hip coordinate
H_r	Right hip coordinate
D_m	Distance ratio from middle shoulder to middle hip relative to shoulder size
D_l	Distance ratio from middle hip to left hip relative to shoulder size
D_r	Distance ratio from middle hip to right hip relative to shoulder size

Table 2. Fixed distance ratios relative to shoulder size

Ratio	Starting point	Ending point	Value
D_m	Middle shoulder (S_m)	Middle hip (H_m)	[0.008628, 1.503246]
D_l	Middle hip (H_m)	Left hip (H_l)	[0.297260, 0.000245]
D_r	Middle hip (H_m)	Right hip (H_r)	[-0.305702, 0.000097]

were subsequently applied to ensure accurate estimation of the missing hip coordinates.

This approach enables the completion of the final step without direct hip point data, thereby ensuring the robustness and adaptability of the overall process.

The subsequent challenge pertains to the variability in body proportions among individuals performing sign language in the input videos. This inconsistency hinders the ability of any virtual character to accurately replicate the movements of these individuals. In this research, we utilize SMPLify-X [32] to generate 3D skeleton motion from 2D motion data. This approach standardizes the motion data onto the SMPL-X 3D model, thereby effectively mitigating the issue of inconsistent body proportions. By addressing this challenge, we also enhance the scalability of the 3D skeleton motion dataset, facilitating the estimation of 3D skeletons from videos produced by a diverse range of individuals.

As SMPLify-X operates as a command-line application, input parameters must be manually specified. To ensure seamless integration into our pipeline, we modified SMPLify-X to develop a module that automates the computation of key parameters, including character type, weight, and display mode, enabling more efficient processing and adaptation to our framework. Additionally, lower body joints (points 10, 11, 13, 14, 19, 20, 21, 22, 23, 24, as illustrated in Fig. 3) are excluded, as sign language datasets typically focus on the upper body, given that hand movements and facial expressions are the primary components of sign communication. Furthermore, due to discrepancies in skeletal control parameter formats between SMPLify-X and the SMPL-X model, the system has been adapted to extract 3D motion data in the required format for sign language animation. For each frame in the videos, we extract the following information:

- **global_orient**: The overall rotation angle.
- **body_pose**: The rotation angle of the body bones (excluding hand bones and the facial region).
- **left_hand_pose**: The rotation angle of the bones in the left hand.
- **right_hand_pose**: The rotation angle of the bones in the right hand.

Subsequently, the data from each frame is aggregated to construct a 3D skeleton motion for a word or phrase.

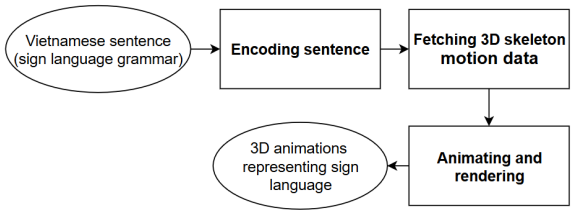

Figure 5. Overview pipeline for generating 3D sign language animation.

3.2. Generating 3D Animations Representing Sign Language

To address the representation of sign language from text in sign language grammar, a process consisting of three steps is proposed, as illustrated in Fig. 5. The input to this process is text that has already undergone word and phrase segmentation, where compound words are denoted by joining their syllables using underscores instead of spaces. The input sentence first is encoded into a format that facilitates efficient retrieval within the dataset. The next step involves searching for these codes within the 3D skeleton motion dataset; any words not found in the dataset will necessitate an execution of the 3D skeleton motion construction process. Once sufficient motion data for the required words has been collected, the final step processes and connects these motions and integrates them into a 3D virtual character to create an animation that represents sign language. The output of this entire process can be exported as a video featuring a virtual character conveying the input sentence in sign language.

3.2.1. Encoding Sentence

The primary objective of this preprocessing step is to identify the sign language symbols corresponding to the words in the input sentence. The process commences by separating the words from the sentence. In the input sentence, compound words are denoted by joining their syllables using underscores instead of spaces. This notation enables a clear distinction of words, facilitating their separation from the sentence based on spaces. Subsequently, we convert all uppercase characters to lowercase and replace the underscore with a space in all words, obtaining a list of words W representing the words and their order in the input sentence. By employing a dictionary $f: X \rightarrow Y$, the list of signs $S = f(W)$ that the virtual character needs to display is retrieved. Words not found in the dictionary will be skipped, and no sign will be displayed for those words.

The dictionary is designed as an injective mapping $f: X \rightarrow Y$, where X is the set of words and Y is the set of signs. In this study, the dictionary is based on the QIPEDC 2D sign language video dataset [33], which contains 4000 VSL words and phrases. The dictionary's architecture is intentionally designed to be expandable, allowing for seamless supplementation with data from other sources in the future. This flexibility ensures adaptability to the evolving needs of the sign language community, accommodating new words and expanding its linguistic coverage over time.

3.2.2. Fetching 3D Skeleton Motion Data

After encoding the words into VSL signs, the corresponding motion data for the words in the input sentence is fetched from the 3D skeleton motion dataset. This data serves as the foundation for animating and controlling the movements of the virtual character. If the required sign data is not available in the 3D sign language dataset, relevant 2D sign language videos will be automatically fetched. Subsequently, these videos undergo a process to generate the necessary 3D skeleton motion data with the proposed method in Section 3.1. The newly created 3D sign language data is then stored in the existing 3D sign language dataset to expand its scope and enhance future performance. This approach not only ensures the availability of diverse sign data but also facilitates the continuous expansion of the dataset, allowing

for the integration of additional available sign language resources.

3.2.3. Animating and Rendering

The input for this step consists of motion data corresponding to the words or phrases in the input sentence. The output is an animation depicting a 3D virtual character representing sign language. It is essential to ensure the coherence of the movements that convey the sentence, as the motion data represents individual words while a sentence is composed of multiple words and phrases. To achieve this, the 3D skeleton motion data must be interconnected and integrated into a virtual character to execute the movements.

To create animations, various techniques can be employed to animate the virtual character. In this study, we utilize Blender as the software for animation construction. A new scene is initialized within the software, comprising two objects: a 3D SMPL-X character and a virtual camera. The virtual character is positioned at the origin $Oxyz$ of the scene's coordinate space. The camera is oriented towards the character's shoulder, maintaining a distance that focuses on the character's hand movements in the rendered images. The camera is set to an Euler angle of $(\frac{\pi}{2}, 0, 0)$.

After setting up the scene, lighting, virtual character, and camera, the virtual character performs each sign corresponding to each word in the sentence. In some frames, the character has fixed poses based on the 3D sign language data, while in other frames, the character's movements are interpolated to create smooth transitions between the fixed frames. The character's movements start in a rest pose in the first frame. Next, the character sequentially adopts the 3D sign language data for each sign identified. Finally, the character's movements end with a return to the initial rest pose. To provide a realistic experience for the user, the first sign's frame appears after about 0.66 seconds from the start of the animation, subsequent signs are displayed approximately 0.33 seconds apart, and the character returns to the rest state within 0.66 seconds after performing the last sign. These specific durations were derived through iterative experiments and consultation with sign language experts to ensure an optimal balance between visual realism, user comfort, and the natural flow of signing. However, motion data when applied to the virtual character has two issues.

The first issue pertains to the rotation angle of the pelvis, which can result in the virtual character appearing inverted. To address this, we propose recalculating the rotation angle using the following formulas:

$$x' = \begin{cases} x + \pi & \text{if } x \leq 0 \\ x - \pi & \text{if } x > 0 \end{cases} \tag{9}$$

$$y' = -y \tag{10}$$

$$z' = -z \tag{11}$$

where x, y, z represent the coefficients corresponding to the Euler rotation of the pelvis bone as controlled by the 3D sign data, while x', y', z' denote the new Euler rotation coefficients following the adjustment.

The second issue pertains to the uncontrolled movement of certain body parts of the virtual character. To address this, the body parts, excluding the hands, are rendered static across all frames, as sign language predominantly emphasizes hand movements for expression, rendering motion in other body parts superfluous. To implement this solution, the rotation angles of all bones governing the legs and head are reset to their default values of $(0, 0, 0)$ in Euler rotation mode following each instance in which the virtual character receives motion data. This methodology ensures that the animation remains concentrated on the pertinent gestures while preserving a consistent and natural appearance. After establishing the virtual character's pose for each frame, keyframes are assigned to each bone to fix the rotation angles at those specific times, thereby enabling the graphics tool to compute the

Figure 6. Virtual character representing the word "đồ ăn"(food) in sign language.

motion. Finally, the entire 3D animation of the virtual character performing sign language is displayed to the user through a video recording of the entire process, from the virtual character starting the first sign to completing the last sign. The video is in mp4 format, with a resolution of 1280x720, and shows 30 frames per second. Fig. 6 captures a sign language performed by the virtual character.

4. Experiments and Results

4.1. Data Descriptions

Using the proposed pipeline, we constructed a 3D skeleton motion dataset for VSL based on the QIPEDC dataset [33]. The resulting dataset consists of over 500 3D skeleton motion samples, each corresponding to a Vietnamese word or phrase. By processing the 2D videos, essential movements were extracted while redundant actions were removed, ensuring that the 3D motion data accurately represents the original gestures. The dataset is continuously being expanded to include more words and phrases, enhancing its coverage and applicability.

The QIPEDC project developed a comprehensive 2D dataset comprising 4,000 VSL signs containing a large number of sign language 2D videos that cover a wide range of vocabulary, including nouns, verbs, adjectives, phrases, numbers, and letters. The primary objective of the project is to improve access for primary-level deaf students through the use of VSL, thereby enhancing their academic performance. The data was collected through video recordings of deaf individuals who possess expertise in VSL, ensuring that the signs accurately represent the language used within the deaf community. This approach not only guarantees the authenticity of the signs but also empowers deaf professionals by actively involving them in the creation of educational resources.

For the purpose of evaluating the effectiveness of the proposed method, we leveraged the QIPEDC dataset to build a 3D skeleton motion dataset for VSL. To date, our dataset has grown to include more than 500 words and phrases, covering a diverse range of categories such as nouns, verbs, numbers, and letters. It continues to be expanded, adding new VSL words and phrases to further enhance comprehensiveness and diversity.

4.2. Experiments

To evaluate the effectiveness of the proposed method, we conducted two experiments designed to address two distinct scenarios: the representation of individual words and the representation of sentences comprising multiple words. Three sign language experts employed at educational institutions for children with special needs in Vietnam participated in the evaluation of the proposed method. The participants viewed videos rendered from 3D animations that depicted the virtual character performing sign language and subsequently responded to a series of provided questions.

The objective of the first experiment is to evaluate the quality of words or phrases represented by the virtual character based on two criteria: semantic accuracy and the quality of movement display. In this experiment, participants viewed videos of individual signs and recorded their meanings, along with scores assessing the display quality. A total of 16 signs were selected for evaluation, consisting of 8 one-handed signs and 8 two-handed signs. Each sign was presented through a video featuring a

virtual character. The selected signs encompass commonly used words in communicative and educational contexts, including nouns, verbs, adjectives, numbers, and letters. This selection ensures that the dataset captures various linguistic categories relevant to sign language communication.

As the experiment sessions were conducted independently, the test dataset for each participant remained consistent in terms of the displayed signs and the order of videos. The semantic responses were categorized as "understand", "not understand", or "misunderstand", which were then used to evaluate the semantic criterion—specifically, the clarity of the signs performed by the virtual character. Participants rated the display quality of the animation on a scale from 1 to 5, where 1 indicated "very poor", 2 indicated "poor", 3 indicated "average", 4 indicated "good", and 5 indicated "very good". The evaluations provided by three experts were aggregated and averaged using the Mean Opinion Score (MOS) method [34].

In the second experiment, participants watched three videos featuring a virtual character performing sign language sentences. They were tasked with interpreting the meaning of each performed sentence and identifying the number of signs it contained. The experiment was designed to assess the virtual character's ability to convey complete sentences rather than isolated words or phrases. The tested sentences incorporated both one-handed and two-handed signs, reflecting the natural variation observed in sign language communication. This approach allowed us to evaluate not only the accuracy of individual sign representations but also the overall coherence and fluency of the virtual character's signing. Following their reports, the responses were evaluated using the Mean Squared Error (MSE) metric [35]. This assessment allowed for an evaluation of the virtual character's proficiency in executing multiple signs within a sentence.

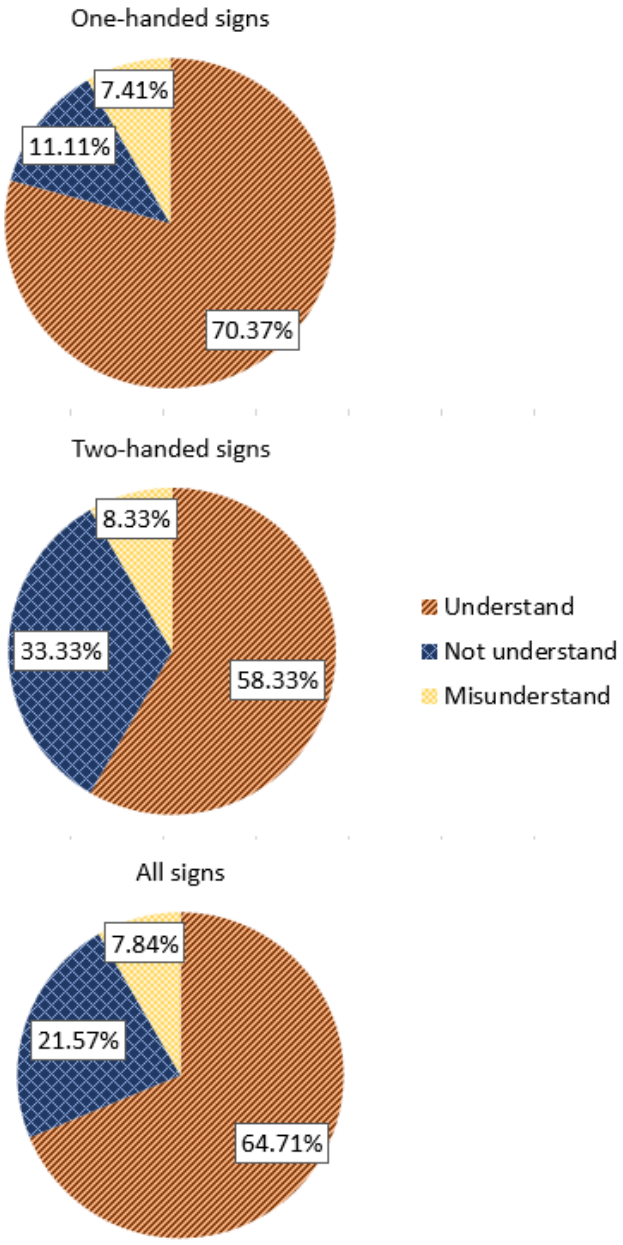

Figure 7. Understanding proportion in words and phrases.

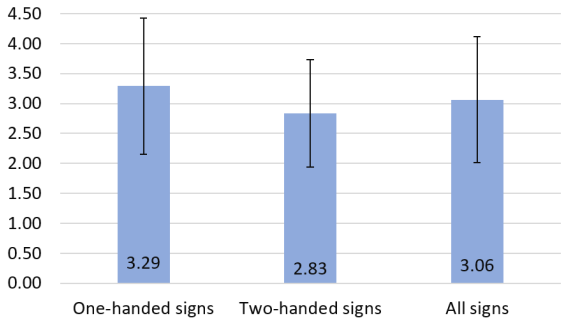

Figure 8. Displaying quality of words and phrases.

4.3. Results

Fig. 7 and Fig. 8 summarize the responses from experts regarding 16 signs, comprising 8 one-handed signs and 8 two-handed signs, represented by virtual characters in the first experiment. The distribution of the symbols, as evaluated by the experimental participants based on semantic assessment criteria, is illustrated in Fig. 7. Additionally, the evaluation results concerning the display quality of each sign are depicted in Fig. 8

The results indicate that the comprehension rate for the generated signs representing sign language is 64.71%. This finding is particularly noteworthy, especially in light of the typical comprehension rates in sign language communication, which generally range from 70% to 75%, as reported by sign language experts. Within this context, our proposed method demonstrates exceptional efficacy in representing single-handed signs, achieving a comprehension rate of up to 70.37%. These results indicate that the proposed method effectively represents sign language, particularly with regard to one-handed signs.

The evaluation results provided by the experts regarding the quality of the rendered videos were aggregated and utilized to calculate the mean and standard deviation, as presented in Fig. 8, according to the equation 12.

$$MOS = \frac{1}{N}\sum_{n=1}^{N} R_n \tag{12}$$

where N is the number of participants and R is the quality score given by the participants. As we can see from the table, one-handed signs received a higher average rating of 3.29, indicating they are perceived as more effective, though their high standard deviation of 1.14 suggests varied opinions among users. In contrast, two-handed signs had a lower average rating of 2.83, with a more consistent rating reflected in a standard deviation of 0.90, suggesting general agreement on their lesser effectiveness. The overall average for all signs was 3.06, indicating room for improvement across the board. These insights suggest a focus on enhancing one-handed signs while re-evaluating the design and usability of two-handed signs to improve user engagement and understanding.

In the second experiment, the results revealed that all experts were able to understand the meaning of every sentence presented with a quite high degree of clarity. Regarding the identification of the number of signs in each sentence, the results are summarized in Table 3. The majority of expert evaluations aligned with the actual number of signs, with only one minor deviation observed in Sentence 3, where Expert 1 reported a different count. The MSE between the average number of signs identified by the experts (X) and the actual number of signs present in the sentences (Y) is calculated using the formula 13.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \tag{13}$$

The MSE value of 0.037 demonstrates the proposed method's ability to represent entire sentences in sign language, rather than being limited to individual, discrete words.

The results of these experiments indicate that the proposed sign language representation method can effectively convey the content of

Table 3. Result of Experiment 2

Sentence	Number of signs	Expert 1	Expert 2	Expert 3
1	7	7	7	7
2	8	8	8	8
3	9	8	9	9

various signs. Notably, the virtual character's ability to perform basic one-handed signs received high ratings, with comprehension levels approaching real-world rates. Additionally, the results of Experiment 2 demonstrate the virtual character's stable ability to perform multiple consecutive signs within a sentence. All the figures confirm the potential of the proposed method outlined in the paper when applied to communication support systems for the hearing impaired.

5. Conclusion

In this study, we have proposed a novel method for representing sign language using 3D virtual characters. The approach involves two stages: constructing 3D skeleton motion for words and phrases, and creating corresponding 3D animations. Unlike existing methods that utilize various forms of notation to represent sign language, our approach introduces a completely new method, marking the first research effort to leverage skeleton motion data for animating characters that represent sign language. Its key advantage is the ease of expanding the vocabulary dataset, as generating 3D skeleton motion data requires only video input showing an individual performing gestures for a single word or phrase. In this study, a dataset comprising over 500 words and phrases in VSL has been constructed using the proposed method and is continuously being expanded.

The output animations were evaluated by sign language experts in Vietnam, yielding positive results. Experimental findings demonstrate that the proposed method effectively represents both individual words and complete sentences in sign language, highlighting its potential for facilitating the translation of text into sign language and advancing accessibility for the deaf community.

Acknowledgment

This work has been supported by VNU University of Engineering and Technology under project number CN24.11.

References

[1] World Health Organization, *Deafness and Hearing Loss*, Accessed: 2024-08-19, Feb. 2024, URL: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[2] Y. Bouzid and M. Jemni, "An Avatar Based Approach for Automatically Interpreting a Sign Language Notation", in: *2013 IEEE 13th International Conference on Advanced Learning Technologies*, 2013, pp. 92–94, https://doi.org/10.1109/ICALT.2013.31.

[3] A. Othman and M. Jemni, "An XML-Gloss Annotation System for Sign Language Processing", in: *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, 2017, pp. 1–7, https://doi.org/10.1109/ICTA.2017.8336054.

[4] B. D. Patel et al., "ES2ISL: An Advancement in Speech to Sign Language Translation using 3D Avatar Animator", in: *2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2020, pp. 1–5, https://doi.org/10.1109/CCECE47787.2020.9255783.

[5] K. Kaur and P. Kumar, "HamNoSys to SiGML Conversion System for Sign Language Automation", in: *Procedia Computer Science* 89 (2016), Twelfth International Conference on Communication Networks, ICCN 2016, August 19– 21, 2016, Bangalore, India, pp. 794–803, https://doi.org/10.1016/j.procs.2016.06.063.

[6] M. Sanaullah et al., "A Real-Time Automatic Translation of Text to Sign Language", in: *Computers, Materials and Continua* 70 (Sept. 2021), pp. 2471–2488, https://doi.org/10.32604/cmc.2022.019420.

[7] D. A. Gonçalves, E. Todt, and L. S. Garcia, "3D Avatar for Automatic Synthesis of Signs in Sign Languages", in: 2015, URL: https://api.semanticscholar.org/CorpusID:64565230.

[8] M. Punchimudiyanse and R. G. N. Meegama, "3D Signing Avatar for Sinhala Sign language", in: *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, 2015, pp. 290–295, https://doi.org/10.1109/ICIINFS.2015.7399026.

[9] O. Al-Barahamtoshy and H. Al-Barhamtoshy, "Arabic Text-to-Sign (ArTTS) Model from Automatic SR System", in: *Procedia Computer Science* 117 (Dec. 2017), pp. 304–311, https://doi.org/10.1016/j.procs.2017.10.122.

[10] H. Luqman and S. A. Mahmoud, "Automatic Translation of Arabic Text-to-Arabic Sign Language", in: *Universal Access in the Information Society* 18.4 (2019), pp. 939–951, https://doi.org/10.1007/s10209-018-0622-8.

[11] M. Brour and A. Benabbou, "ATLASLang MTS 1: Arabic Text Language into Arabic Sign Language Machine Translation System", in: *Procedia Computer Science* 148 (Jan. 2019), pp. 236–245, https://doi.org/10.1016/j.procs.2019.01.066.

[12] Q. P. Van and B. N. Thanh, "Vietnamese Sign Language Recognition using Dynamic Object Extraction and Deep Learning", in: *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*, 2021, pp. 402–407, https://doi.org/10.1109/ICCE48956.2021.9352071.

[13] B. Duy Khuat et al., "Vietnamese sign language detection using Mediapipe", in: *Proceedings of the 2021 10th International Conference on Software and Computer Applications*, ICSCA '21, Kuala Lumpur, Malaysia: Association for Computing Machinery, 2021, 162–165, https://doi.org/10.1145/3457784.3457810.

[14] L. D. Quach and C.-N. Nguyen, "Converting the Vietnamese Television News into 3D Sign Language Animations for the Deaf", in: *Lecture Notes of the Institute for Computer Sciences* 257 (Jan. 2019), https://doi.org/10.1007/978-3-030-05873-9.

[15] O. ElGhoul and M. Jemni, "WebSign: A System to Make and Interpret Signs Using 3D Avatars", in: *Proceedings of the Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT), Dundee, UK*, vol. 23, 2011.

[16] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs", in: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 723–732.

[17] L. Ma et al., "Pose guided person image generation", in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, 405–415.

[18] R. Rastgoo et al., "Sign Language Production: A Review", in: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 3446–3456, https://doi.org/10.1109/CVPRW53098.2021.00384.

[19] L. Naert, C. Larboulette, and S. Gibet, "A Survey on the Animation of Signing Avatars: From Sign Representation to Utterance Synthesis", in: *Computers Graphics* 92 (Nov. 2020), https://doi.org/10.1016/j.cag.2020.09.003.

[20] S. Cox et al., "The Development and Evaluation of a Speech-to-Sign Translation System to Assist Transactions", in: *Int. J. Hum. Comput. Interaction* 16 (Oct. 2003), pp. 141–161, https://doi.org/10.1207/S15327590IJHC1602˙02.

[21] F. Pezeshkpour et al., "Development of a Legible Deaf-Signing Virtual Human", in: *Proceedings IEEE International Conference on Multimedia Computing and Systems*, vol. 1, 1999, 333–338 vol.1, https://doi.org/10.1109/MMCS.1999.779226.

[22] S. Alexanderson and J. BESKOW, "Towards Fully Automated Motion Capture of Signs – Development and Evaluation of a Key Word Signing Avatar", in: *ACM Trans. Access. Comput.* 7.2 (June 2015), https://doi.org/10.1145/2764918.

[23] S. Gibet et al., "The SignCom System for Data-Driven Animation of Interactive Virtual Signers: Methodology and Evaluation", in: *ACM Trans. Interact. Intell. Syst.* 1.1 (Oct. 2011), https://doi.org/10.1145/2030365.2030371.

[24] R. Brun, A. Turki, and A. Laville, "A 3D Application to Familiarize Children with Sign Language and Assess the Potential of Avatars and Motion Capture for Learning Movement", in: *Proceedings of the 3rd International Symposium on Movement and Computing*, MOCO '16, Thessaloniki, GA, Greece: Association for Computing Machinery, 2016, https://doi.org/10.1145/2948910.2948917.

[25] M. Jemni and O. Elghoul, "A System to Make Signs Using Collaborative Approach", in: July 2008, pp. 670–677, https://doi.org/10.1007/978-3-540-70540-6˙96.

[26] S. Prillwitz and H. Z. für Deutsche Gebärdensprache und Kommunikation Gehörloser, *HamNoSys: Version 2.0; Hamburg notation system for sign languages; an introductory guide*, Signum-Verlag, 1989.

[27] R. Kennaway, "Avatar-Independent Scripting for Real-Time Gesture Animation", in: (Nov. 2006), https://doi.org/10.13140/2.1.2579.2002.

[28] V. H. Group, "Virtual Humans Research for Sign Language Animation", in: School of Computing Sciences, UEA, Norwich, UK.

[29] D. H. L. Nguyen, H. L. Luong, and T. D. Ngo, "A Pipeline for 3D Skeleton Motion Estimation from 2D Videos for Sign Language Representation", in: *Proceedings of the KSE 2024*, Accepted for publication, 2024.

[30] Z. Cao et al., "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2021), pp. 172–186, https://doi.org/10.1109/TPAMI.2019.2929257.

[31] A. Moryossef et al., *Real-Time Sign Language Detection using Human Pose Estimation*, Aug. 2020, https://doi.org/10.48550/arXiv.2008.04637.

[32] G. Pavlakos et al., "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image", in: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] *The Quality Improvement of Primary Education for Deaf Children Project (QIPEDC)*, Global Partnership for Results-Based Approaches (GPRBA), URL: `https://qipedc.moet.gov.vn/`.

[34] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives", in: *Multimedia Systems* 22.2 (2016), pp. 213–227, https://doi.org/10.1007/s00530-014-0446-1.

[35] D. Chicco, M. Warrens, and G. Jurman, "The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation", in: *PeerJ Computer Science* 7 (July 2021), e623, https://doi.org/10.7717/peerj-cs.623.