Cureus Journal of Computer Science

Open Access Original Article

Development of an Animation-Based Indian Sign Language e-Learning App for Persons With Hearing and Speech Impairment

Review began 07/22/2024
Review ended 11/19/2024
Published 11/24/2024

DOI: https://doi.org/10.7759/s44389-024-00347-8

Navneet Nayan [1] , Debashis Ghosh [1] , Pyari M. Pradhan [1]

1. Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee, Haridwar, IND

Corresponding authors: Navneet Nayan, nnayan@ec.iitr.ac.in, Debashis Ghosh, debashis.ghosh@ece.iitr.ac.in

Abstract

Human-computer interaction (HCI) is changing the quality of our lives. In this paper, we present one such application of HCI that comes as a boon in the life of a hearing-impaired person. We propose to build an e-learning tool for the education of the hearing- and speech-impaired community by creating animated gestures in sign language, with particular emphasis on the Indian Sign Language (ISL). With the help of MATLAB App Designer, we developed a standalone MATLAB application for ISL fingerspelling (including digits, alphabets and combination of these two) and some ISL words of daily usage. This app can be used by the hearing-impaired community for learning ISL, even while seated at home, or for training signers and teachers at special schools. In our proposed method, the software first learns different sign gestures from real gesture videos. For this, hand regions in every frame are extracted using image processing algorithms and subsequently identifying key video frames that show significant change in hand shape and/or position during gesturing. Correlation coefficient and structural similarity are the two metrics used for the purpose. Next, several hand parameters such as the finger joint angles and orientation of the palm in these key frames are derived using Google MediaPipe holistic model. When using the app, for a query sign, the corresponding gesture is animated by synthesizing the key frames containing 3D hand shapes with desired palm orientation and finger joint angles using the stored hand parameters, followed by interpolating the in-between frames using image metamorphosis. For interpolation, we have used spline interpolation and intensity interpolation. Some sign gestures created in our experiments show that our proposed method generates smooth and natural-looking animated videos from a comparatively less amount of stored gesture information, thereby offering large savings in memory.

Categories: Animation, HCI Appliances and applications, Computer Vision
Keywords: gesture synthesis, gesture animation, indian sign language, key frame extraction, hand extraction, hand parameter extraction, image metamorphosis, hand gesture animation

Introduction

With the emergence of low-cost computers and increased popularity of personal computers among common people, efforts are being made to use this resource in our daily lives, including business, education, entertainment, etc. In this context, several e-learning tools have been developed and used successfully in distance learning. Furthermore, with the advancements in network infrastructure and multimedia technology, web-based online teaching has been widely applied in educating people without really attending conventional classroom lectures. Also, the impact of the COVID-19 pandemic resulted in the surge of demand for online education. This led to the development of novel and innovative ideas for online education. However, while much has been done for the development of computer-assisted learning for normal being, at present, very little is available for people who are physically challenged, particularly for the community with hearing and speech impairments. So, it is our social responsibility that a significant amount of research be devoted in utilizing the modern technology to make the lives of deaf easier and facilitate their relationship with other individuals with normal hearing ability.

Nothing is more important for the community with hearing and speech impairments than deaf education. It is a matter of debate whether a child suffering from hearing and speech impairments should be sent to a conventional mainstream school or to a specialized school. While mainstream schools can provide higher level of education, classroom accessibility is in question. People with hearing and speech impairments are often handicapped because information is presented in a form inaccessible to them. On the other hand, sign language-assisted learning alleviates barriers to communication on a larger scale and helps remove the handicap. Sign language is the language of the people suffering from hearing and speech disabilities. It is used for expressing what they want to “say”. Therefore, every person with hearing and speech impairments requires to learn “how to speak” using sign language. There are schools for the purpose. Sign language is used as the medium of instruction in specialized schools providing better access for the children suffering from hearing and speech disabilities. However, such specialized schools are generally located in cities and big towns only. As a result, a major portion of the population with hearing and speech impairments that resides in rural and remote locations is deprived of even the very basic level of education. Consequently, a major portion of the population with hearing and speech impairments is still “illiterate”. Nevertheless, with the recent advances in computer technology, it may be possible to reach some level of education among the community with hearing and speech impairments in rural areas. Utilizing technology to provide sign language as a learning tool, we seek to expand opportunities for communication and provide information access to the community facing hearing and speech impairments, thereby bridging the gap between two worlds: one world comprising persons suffering from hearing and speech disabilities, and the other world consisting of persons having normal hearing and speech abilities. This effort will allow people with hearing and speech disabilities to achieve their potential in educational, vocational and personal pursuits.

Human-computer interaction (HCI) has made our lives easy by providing solutions to many day-to-day problems. In this regard, HCI has played a major role in improving the quality of life of people who are differently abled. When it comes to the community with hearing and speech impairments, last few years have seen a significant involvement of HCI in their education via sign language. This has led to enabling and empowering them in almost every aspect of their lives, thereby providing an opportunity to communicate and compete with the normally abled people of the society. This has been achieved by developing efficient frameworks for sign language recognition and sign language generation. While sign language recognition helps a normal person to understand signs performed by people who are specially abled, sign gesture generation helps a person with hearing impairment to understand the speech spoken by a normal person by translating the speech to corresponding gestures in sign language. Both of these, sign language recognition and sign language generation, together can form a conversation system for exchange of dialogs between a person with hearing impairment and a normal person. This, the conversation system, supports continuous dialog between a person with hearing impairment and a normal person, even without an interpreter.

The sign language synthesis module described above can also be very much used in developing computer-aided e-learning tools for the education of the community with hearing and speech impairments. Deployment of internet even in remote areas and easy availability of smartphones have further boosted the HCI research community to look for developing such e-learning tools.

The primary aim of this work is to develop a computer-assisted learning tool for Indian Sign Language (ISL), thereby facilitating e-learning of sign language for the community in India with hearing and speech impairments so that an individual with hearing and speech impairments can learn sign language while seated at home and even without attending a special school. These days, many websites and mobile apps are available for the purpose of educating the community with hearing and speech impairments through e-learning of sign language. These e-learning tools are mostly dictionary-based, where for every query input text, corresponding gesture video (gesture sequences) is played back. Accordingly, several online dictionaries for some sign languages are available in the internet that are used for home-learning of sign languages [1-3]. Buttussi et al. [1] proposed 3DictSL (an online international sign language dictionary) to simplify sign-to-word conversion and sign-to-sign conversion. For this, they have utilized several Web3D technologies like H-Anim humanoids and X3D. One of the major limitations of the work is related to the visualization of the website. This could be only visualized using some of the recent browsers, with X3D players supporting Javascript SAI and H-Anim nodes. Cavender et al. [2] have developed ASL-STEM forum to develop American Sign Language (ASL) in a virtual environment. This forum is an online and collaborative video forum made to share the signs of ASL and discuss the issues in ASL. In this, community members upload their live videos related to the signs and concepts of ASL that were further streamed on YouTube. This requires a good amount of internet speed to upload and play the video. Han et al. [3] presented a framework to translate signs into English. Breaking the signs into several sub-units and extracting the spatiotemporal features from them are the key points of this research work. On British Sign Language (BSL), the authors obtained improved and satisfactory results. However, to make it easy and accessible for the public, they proposed to develop a portable device for the sign language translation based on their proposed approach. In all of these works, authors use prerecorded digital video clippings corresponding to different signs. However, this demands large amount of memory to store gesture videos for a large vocabulary. Furthermore, most of these dictionaries have been developed for ASL and others for sign languages of European nations like BSL, French Sign Language (FSL), etc. No such system was available for ISL until a multilingual multimedia ISL dictionary-building tool was developed in recent times [4,5]. Nevertheless, their use is restricted due to large memory requirement and limited vocabulary.

A solution to this is to make a computer create the gestures for any desired sign. Compared to gesture recognition, research in gesture synthesis is less explored. In the context of ISL, this lacuna in gesture synthesis research is more pronounced. The main reasons behind this are as follows: (1) lack of awareness in ISL, (2) lack of research in ISL linguistics, (3) unavailability of well-documented and annotated ISL lexicon and (4) availability of very little literature on ISL. Consequently, although similar modules are available for ASL, BSL and FSL, it is missing for ISL. As a result, the community of hearing and speech impairments in India faces difficulty in communication not just with the hearing world but also in peer-to-peer interaction. In view of this, it is important that a technology-based learning tool be developed to impart knowledge of ISL among people in India suffering from hearing and speech impairments. Therefore, the need to build a system that can associate signs with the words of spoken language, and which can further be used to learn ISL, is significant.

While people with hearing and speech impairment will be the primary beneficiary of this proposed e-learning tool, it can as well be used by the community at large that comes in contact with people with hearing and speech impairments. The tool can be used to train academic staff at special schools/integrated schools, and for self-training of sign language interpreters. Sign language is not only used by those with hearing and speech impairments but also by their hearing parents and/or their hearing children for day-to-day communication at home. Parents/children of people with hearing and speech impairment can use this e-learning tool to learn sign language at home so as to facilitate them to communicate easily with their children/parents. Thus, the Indian society as a whole will be benefitted, bridging the gap between the hearing-impaired and the hearing world.

Here in this paper, we focus to develop a mobile application based on synthesis and animation of ISL for the purpose of education of the community in India suffering from hearing and speech impairments. The unique contributions of our proposed tool are as follows: it does not require any image or video database to be stored to play the animation, thereby not requiring any major requirement of system memory, and to the best of our knowledge, in the context of ISL, an animation-based teaching and learning tool for the education of hearing speech community is missing. Therefore, our proposed tool may help fill this gap. As mentioned, our proposed work is focused on ISL gestures that comprise one-handed signs and ambidextrous signs. Two major aspects to be considered in animating sign gestures are smooth transition from one pose to other and maintaining the naturalness during gesturing. Our proposed method addresses both of these aspects and produces natural and smooth animation of sign gestures, and works well for both local and global hand motions. The proposed method is based on using key frames to generate the animation of gesture videos that comparatively requires less memory, thereby mitigating the requirement of large memory.

The key research issues involved in this work are as follows: feature extraction to obtain a suitable set of spatio-temporal features good for quantitative description of dynamic gestures, feature selection and gesture description to create a database of sign gestures with minimum memory requirement, and gesture animation scheme for generating desired sign gestures using stored information. Overall, our contribution of the proposed work can be summarized as follows:

(1) Developing a standalone application for educating the community suffering from hearing and speech impairments in the context of ISL.

(2) Extracting hand region efficiently using our image processing algorithm.

(3) Efficient key frame extraction based on correlation coefficient and structural similarity.

(4) Hand parameter extraction for ISL signs using Google MediaPipe holistic model.

(5) Sign language synthesis and animation of ISL fingerspelling (digits, alphabets and combination of both) and ISL words.

(6) Developing ISL fingerspelling and ISL words dataset following the database of Indian Sign Language Research and Training Centre.

Materials And Methods

Literature survey

Although literature in gesture synthesis is rather thin, some researchers in this area did attempt to build some gesture animation schemes. A Virtual Reality Modeling Language (VRML)-based ASL finger-spelled system is the first reported work of this kind [6-9]. Geitz et al. [6] established a teaching tool that represents manual alphabet hand shapes of ASL. This tool consists of 3D computer graphic models of ASL. These models are recorded in the VRML. For displaying these models, the World Wide Web browsers were used. Next, a speech signal to ASL translation system is proposed by Patel and Rao [7]. This system contains the pre-recorded signs that are displayed corresponding to the input words. If signs corresponding to the words are not present, then those words are finger-spelled. The only issue with this system is the memory requirement to store the pre-recorded signs. Seki and Sato [8] used head-related transfer function simulation to develop a new auditory orientation training system. They also used acoustic virtual reality to develop this system and obtained satisfactory results. A multimodal system to develop an educational and entertainment tool is explained by Moustakas et al. [9]. This tool helps hearing and visually impaired community to learn and get education. Speech, sign language and haptics are different modalities used in this work.

In the sequence of these research works, some other researchers developed sign language animation schemes [10-13], with applications to embodied conversational agents[14-18], and speech/text to sign language translation systems [19-24]. Sedgwick et al. [10] developed a system for ASL fingerspelling animations. They developed hand models and used interpolation to complete the task of animation. They obtained high recognition rates by deaf users. However, users reported confusion in animated signs of some of the alphabets. Kennaway [11] discussed a method for the synthesis of deaf signing animation system. To develop this system, the author used HamNoSys transcription system to describe the signs. This method performs the automatic synthesis of signing gestures based on their description in HamNoSys notation. Verlinden et al. [12] have discussed the integration of multimedia and animated sign language to develop the teaching material for hearing and speech impaired community. To create the animation of sign language, special sign language notation system was used. The authors developed several web applications and asked the deaf users to determine the efficacy of their proposed method. They obtained good feedback about their proposed method. However, they also focused on adding some more technicalities to their system as future work. Söylev and Mendi [13] proposed data-driven and real-time Turkish Sign Language animation system. They used motion capture techniques to serve their purpose of animation. They reported that optical systems provided more accurate results compared to Microsoft Kinect sensor. Kopp and Wachsmuth [14] reported a novel framework for the generation of speech and gesture from XML-based description. The proposed model focused on synthesizing synchronized verbal, gestural and facial behaviors. To make the synthesis fluent and natural, transition effects and co-articulation were carefully addressed. In this work, the authors presented a kinematic approach for efficient animation of hand gestures. Olivier [15] addressed the issue of generation of gestures and their co-ordination relevant to the spoken words or utterances. The author worked on developing spontaneous gesture synthesis for embodied conversation agents (ECAs). For this, the author utilized the relation between the linguistic properties of the spoken words and graphical model of the characters. In similar sort of work, Hartmann et al. [16] synthesized expressive agent gestures by focusing on creation of an expressive ECA. For this, they presented a computational model of gesture quality. In the proposed ECA system, the authors demonstrated animations under various expressivity settings. Depending on the user reviews, they reported satisfactory results using their proposed approach. In a work on developing ASL animation system, Huenerfauth et al. [17] reported a prototype of ASL generation component. They implemented and evaluated their prototype and found major challenges. The evaluation done in this case also is user-based evaluation. Wik and Hjalmarsson [18] used two systems, namely Ville and DEAL, for the language learning purpose. Both of these systems used ECAs. Ville is used for pronunciation and vocabulary learning, whereas DEAL is a dialogue system to practice conversational skills.

An English language to sign language translation system is presented by Sáfár and Marshall[19]. They implemented natural language processing (NLP) techniques to generate semantic representation from the English text. The semantic representation was based on discourse representation structure, and the NLP implemented was based on semantic, syntactic and discourse-oriented information. In another work on gesture generation, San Segundo et al. [20] generated gestures from speech. The proposed work used Spanish Sign Language (LSE) for their testing. The proposed approach consisted of four basic modules, namely, speech recognition module, module for semantic analysis, gesture generation module and module to play the gestures. Interpolation technique is used in the gesture generation module. They obtained satisfactory outcomes using their proposed approach. Another text to sign language conversion system is presented by Sarkar et al. [21]. This system is made in the context of Bangla language. They proposed a translator software to serve their purpose and obtained encouraging results. In this sequence of text to sign language conversion, Almohimeed et al. [22] proposed a system for the same in the context of Arabic Sign Language. On a corpus of 203 sign sentences (comprising 710 distinct signs), the authors obtained 46.7% and 29.4% of average word error rate and average position error rate, respectively. Another machine translation system that translates Spanish into LSE is proposed by San-Segundo et al. [23]. The proposed system contained three basic modules, namely, speech recognition module, module for natural language translation and a 3D avatar animation module. The developed system was made to help deaf people in renewing their driving license. A field evaluation was done to test the efficacy of the proposed system. The authors obtained highly encouraging results after the evaluation process. In another extended work on the evaluation of machine translation system of LSE, Lopez-Ludena et al. [24] proposed a novel framework for this. They also developed an advanced speech communication system for the deaf people. The proposed system contains two modules, namely, Spanish to LSE translation module and spoken Spanish generator module. After performing the evaluation part, they reported satisfactory results and also some major problems of the proposed system. Table *1* briefly summarizes some of the major works discussed here. Looking at all these works, we find that some complete systems for gesture/sign generation are commercially available in some other countries. Examples are Visicast system, Zardoz system, SiSi system, etc. However, all these available systems are based on ASL, BSL, FSL, LSE, etc., and, hence, are not useful in the Indian context. Accordingly, a system for ISL synthesis is much desired.

Research Paper ID	Major Contribution of the Research Paper
[6]	Teaching Tool for ASL
[7]	Speech Signal to ASL Translation System
[9]	Multimodal-Based Educational Tool for Hearing-Impaired Community
[10]	ASL Fingerspelling Animations
[11]	HamNoSys-Based Synthesis of Deaf Signing System
[12]	Integration of Animated Sign Language into Multimedia
[13]	Data-Driven and Real-Time-Based Turkish Sign Language Animation System
[14]	XML-Based Generation of Gesture and Speech
[15]	Spontaneous Gesture Synthesis for ECA
[16]	Creation of an Expressive ECA
[17]	Prototype of ASL Generation Component
[18]	Development of Language Learning Systems Using ECAs
[19]	English Language to Sign Language Translation System
[20]	Gesture Generation from Speech in the Context of LSE
[21]	Text to Sign Language Conversion for Bangla Language
[22]	Text to Sign Language Conversion for Arabic Language
[23,24]	Machine Translation System for Conversion of Spanish to LSE

TABLE 1: Major Contribution of Some of the Earlier Reported Works

Some of the early works on gesture or sign language synthesis in the context of our proposed method have been reported in various studies [25-27]. Verma and Ghosh [25,26] used spheres, cones and cylinders for developing the hand models for the purpose of gesture synthesis by animation. In the entire process of animation, they performed key frame extraction followed by calculating the hand parameters (hand orientation and finger joint angles) and subsequent reconstruction of the 3D hand shapes using the corresponding hand parameters. However, both these works focused only on gesture animation of single-handed gestures. Also, smoothness and naturalness were missing in the generated animated videos due to the hand shapes formed using regular 3D figures. Working on improving naturalness, Shankar and Ghosh [27] produced a gesture synthesis and animation method for dynamic hand gestures using image morphing. They obtained satisfactory results on one-handed dynamic gestures with improved smoothness and more natural-looking gestures. In this work also, the authors first performed key frame extraction, followed by hand parameter calculation. Next, the obtained hand parameters were used for reconstructing the key frames with differently shaped life-like synthetic hands. Finally, interpolation was performed to generate frames between two key frames to create a continuous animated gesture video. Introduction of image metamorphosis during interpolation led to a smooth transitioning between frames, thereby improving naturalness in animated gestures. However, they also used their approach in animating single-handed gestures only. Another work discussing the animation of single-handed gestures is by Bhuyan et al. [28]. In this work, they created animation of single-handed gestures, but with two major limitations. One is the use of only frontal postures to calculate the hand parameters. The second one is not perfectly replicating those gestures that had partially bent fingers. Another major work focusing on single-handed and ambidextrous hand gesture animation is given by Bhuyan et al. [29]. Here the authors reported a more natural hand gesture animation that alleviated the limitations of the earlier works. They developed a scheme for animating local hand motions and proposed to extend this work for the animation of global hand motions as their future work. A research work focusing on developing an animation system for generating ISL gestures is mentioned by Kaur and Singh [30]. In this, the authors presented a system that generated Hamburg Notation System (HamNoSys) for ISL words. The authors used 100 ISL words of daily use for their work and generated HamNoSys for these words. They covered both one-handed signs and also the ambidextrous signs. They verified their accuracy with the help of ISL dictionary and reported encouraging results. However, the proposed method did not target complex ISL words and lacked implementation on large-scale ISL dataset. Also, animation of ISL alphabets, digits and ISL fingerspelling are missing in this work. As future prospects, the authors showed interest in working towards these lacunae in their work.

Last few years have seen the emergence of application of deep neural networks in image generation and animation that encompasses synthesis and animation of faces, hands and even the entire body. Introduction of deep architectures has led to very smooth and natural animation. However, this comes at the cost of increased computation. Some of these works are discussed in various studies [31-34]. In one of the early breakthroughs of image animation, Wiles et al. [31] presented a novel and robust self-supervised neural network model named as X2Face, which worked without requiring any prior knowledge of pose, expression and input image identity. The proposed architecture is used for generating frames by controlling a source frame using a driving frame to produce a generated frame with the identity of the source frame. During the frame generation process, the network uses the identity of the input source frame and extracts information about pose and expression from the driving video frames. The authors claimed that for further generation process driven by other modalities like pose codes and audios, there is no requirement of additional training of the network. Compared to other self-supervised image generation methods, the proposed network is more robust and versatile, but lacks in producing high-quality synthetic images. In a work on gesture-to-gesture translation, Liu et al. [32] proposed a novel method for image-to-image translation using simple user-friendly annotations without requiring skeleton labels or key points, thereby reducing the annotation efforts. They proposed a novel generative adversarial network architecture that helped in specifying the gesture type as well as provided geometrical and structural details of the hand gesture. Use of rolling guidance approach with an attention module helped in generating high-quality images. They tested their approach on two hand gesture datasets, namely, the NTU Hand Gesture dataset and the Creative Senz3D dataset. Here, the authors obtained state-of-the-art results and surpassed the previously reported state-of-the-art works both in quality and quantity.

Toledano et al. [33] presented a novel idea for video frames synthesis of a source image according to the motion obtained from a driving video. Broadly, their proposed network may be segmented into two parts: one obtaining the masks and the other responsible for frame synthesis. In the first segment, the authors generated two versions of mask. One is key point heatmap mask, while the second one being circle mask. The authors performed their experiments on Tai-chi dataset and obtained comparable results claiming their method to be more feasible but at the cost of some artifacts. Tao et al. [34] introduced a new method, namely, the motion transformer, for image animation. To predict the motion information from a driving video, all previously existing CNNs model the motion interactions explicitly resulting in apparent artifacts in the generated animation videos. This problem led to the development of the motion transformer. It is based on the concept of vision transformer that estimates motion from the driving video successfully by predicting the key points and local transformations. In this paper, the authors tested their proposed method on VoxCeleb, Tai-Chi-HD, TED Talks and MGif datasets and claimed to obtain promising results by effectively capturing the global motion information. An important application of image animation is virtual try-on. Many such applications have reported satisfactory performances in this case. Nevertheless, these methods are either multi-stage frameworks dealing with body blending and clothes warping or relied heavily on intermediate parser-based labels that might be inaccurate or noisy.

Besides these synthesis and animation works, another important aspect in animation is the extraction of hand parameters or hand key points. Some earlier works involve estimating local hand motion via extraction and frame-to-frame tracking of certain key points in the palm region such as the fingertips, metacarpophalangeal (MCP) joints, proximal interphalangeal (PIP) joints and distal interphalangeal (DIP) joints. With the advancement of deep neural networks, several other robust models have been developed that extract hand key points accurately. One of these models is discussed by Simon et al. [35] and the other one is MediaPipe Holistic model developed by Google [36]. Simon et al. [35] used multiview bootstrapping procedure for detecting the hand key points. This work is reported as the first real-time hand key point detector in wild and complex red, green and blue (RGB) images and videos. Google MediaPipe holistic model [36] for landmark detection detects 21 hand landmarks per hand, as depicted in Figure*1*. This model is also highly robust and accurate.

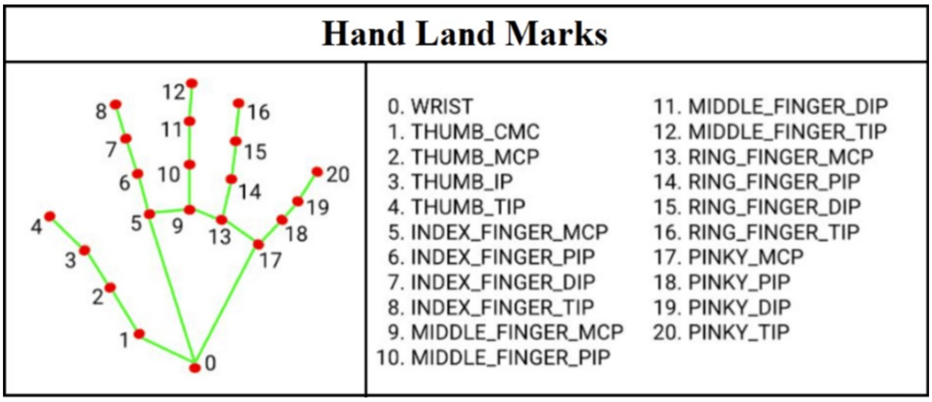


FIGURE 1: Model showing 21 hand landmarks per hand that Google MediaPipe holistic model develops.

Proposed method

Our proposed method of sign language animation proceeds through the following steps:

1. Hand region extraction from input sign gesture video.

2. Key frame extraction based on the similarity of hand postures.

3. Hand parameter estimation.

4. Sign gesture animation.

The first three steps involve learning the gesture sign to be animated. Required frame-to-frame movement of the hands for generating animated gesture sign is learned from real gesture video in which an expert signer do the gesturing. Local hand motion is learned by capturing frame-to-frame change in orientation of the palm and finger joint angles. Global hand motion is learned by tracking the movement of the arm from one frame to the next. Finally, automatic animation of synthesized realistic gesture is accomplished by translating the learned gesture information into an appropriate sequence of hand poses.

Hand Region Extraction

In the first step, the hand region is extracted from an RGB image (each frame of the gesture video) using the following steps:

1. Face detection using Viola-Jones algorithm [37]. For the purpose of face detection, a cascade object detector based on Viola-Jones algorithm is prepared and used to detect the faces. This algorithm uses the AdaBoost learning procedure to make the system learn from the obtained facial features. This learning process is followed by cascading of classifiers to detect and classify the faces. Viola-Jones algorithm is highly accurate in detecting the frontal faces that is required in our scenario of face detection of a signer. In our case, the signer always faces the camera while performing a sign. So, Viola-Jones worked accurately to detect faces in our scenario.

2. Removal of face portion from the image leaving behind the hand and neck portion as the only skin colored regions in the image.

3. Conversion of the image obtained in the previous step from RGB color space to YCrCb color space. YCrCb color space represents color with luminance (Y) and two color difference values Cr and Cb formed by subtracting Y from red and blue components. This color space is less sensitive to ambient light conditions.

4. Detecting skin regions in the image based on image segmentation via K-means clustering for unsupervised classification of similar colored pixels. In our process, image pixels are clustered into three clusters: background (uniformly green colored background taken in our work), signer's clothes (in our work, uniformly dark colored clothes worn by the signer) and skin pixels. Thus, skin pixels are identified, thereby segmenting out the hand and neck region in the image frame.

5. Detection and removal of the neck region. At the end of our previous step, we generally have three

different skin regions corresponding to the two hands and the neck, the smallest sized region being the neck region. In some frames, the two hand regions may also be joined. The neck region is, therefore, easily identified and removed on the basis of the area of extracted skin regions.

6. Finally, the coordinates of the hand pixels are obtained to locate the position of the hands in the image frame.

Algorithm 1: Algorithm to Extract Hand Regions from a Hand Gesture Video

Input: RGB Frames of Hand Gesture Video

Output: Extracted Hand Regions from the RGB Frames

```
Preparing a face detector based on Viola-Jones algorithm
Face detection using the face detector followed by face removal from the input RGB frame. We name this image frame as $Img$.
Conversion of image frame obtained in the previous step from RGB color space to YCrCb color space.
$Converted_{Img}$ = rgb2ycbcr ($Img$);
K-means clustering based Skin Region Segmentation
$Skin_{Img}$= imsegkmeans ($Converted_{Img}$ ,3);
$Hand_{Regions}$ = Detecting and Filtering out the Neck Region based on the Area of Extracted Skin Regions in $Skin_{Img}$.
Obtaining the coordinates of hand pixels in $Hand_{Regions}$ to locate the position of the hands in the image frame $Img$ to produce $Extracted_{Hands}$ .
RETURN Extracted Hand Regions
```

The above procedure for hand region extraction is illustrated in Figure 2.

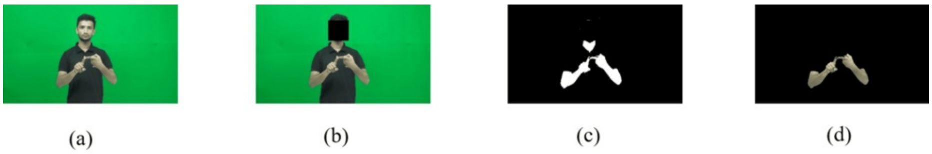


FIGURE 2: An example of hand region extraction from an RGB input frame. (a) Input RGB video frame, (b) frame obtained after face removal, (c) Y-frame with skin regions segmented out, and (d) RGB-frame with hand region segmented out.

Key Frame Extraction

In simple words, key frames are the frames that undergo significant changes in hand pose compared to its immediate preceding key frame. In our proposed method, for key frame extraction, starting with the very first frame in the input gesture video sequence as the first key frame, every successive frame is compared to its preceding key frame. If the difference exceeds some empirically obtained threshold, then the candidate frame is marked as the next key frame in the sequence. The process is repeated with this new key frame till the last frame of the sequence. The last frame is marked as the default last key frame in the sequence. In our work, we used correlation co-efficient and structural similarity metric to compare the candidate frame against its preceding key frame. Thus, the key frames in the input gesture video are extracted. The entire long sequence of the video may now be represented in a compact form by this limited set of key frames, while discarding all in-between redundant frames that hardly show any temporal change. Figure 3 shows the key frames extracted from a gesture sequence for the ISL alphabet W (that contained more than 100 frames) and demonstrates how a large gesture sequence is summarized into a few numbers of key frames preserving all necessary information regarding the movement performed by the hands while gesturing the corresponding sign. These key frames, thus, form the anchor points to create an action in an animation sequence.

Algorithm 2: Algorithm for Key Frame Extraction from an Input Gesture Video

Input: Frames of Input Gesture Video

Output: Extracted Key Frames of the Input Gesture Video

```
Initialization:
Reading Input Gesture Video as $InGv$
Obtaining the total number of frames of $InGv$ and naming it as $TN_{Frame}$
Making the very first frame of $InGv$ as first key frame. We name it $KF_1$.
Reading frames of the current video successively.
Comparing every successive frames to its preceding key frame based on correlation co-efficient and structural similarity
for t =1 : 1: $TN_{Frame}$
    $InGv_{frame}$= read($InGv, t$);
    $InGv_{GrayFrame}$ = rgb2gray ($InGv_{frame}$);
    $Struct_{Sim}$ = ssim ($InGv_{frame}$, $KF_1$);
    $Corr_{Coeff}$ = corr2($InGv_{GrayFrame}$, (rgb2gray($KF_1$));
    if $Corr_{Coeff}$ <Threshold$_1$ & $Struct_{Sim}$ < Threshold$_2$
        Make $InGv_{frame}$ as the Next Key Frame $KF$
        $KF_1 = KF;$
    else
        Continue Looking for the Next Key Frame
    end
end
RETURN Extracted Key Frames
```

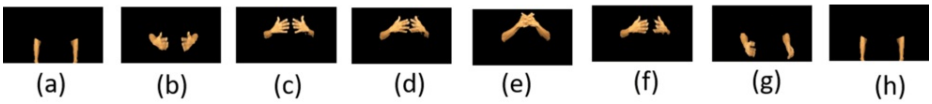


FIGURE 3: Extracted key frames from the gesture sequence of the ISL alphabet W.

Hand Parameter Estimation

For hand parameter estimation, we used Google MediaPipe Holistic model. This MediaPipe hand landmarker helps to find and locate important hand landmarks in an image. Various hand key points detected by this model are wrist joint; thumb carpometacarpal joint; thumb interphalangeal joint; thumb MCP joint; tip of the thumb; MCP, PIP and DIP joints; and fingertips of other fingers. This model works equally well for both single-handed and double-handed gestures. As an example, in Figure 4, we show the hand landmarks detected in the image in Figure 2(d). With the help of MediaPipe model, we were able to extract the hand landmarks in image coordinates as well as in world (global) coordinates.

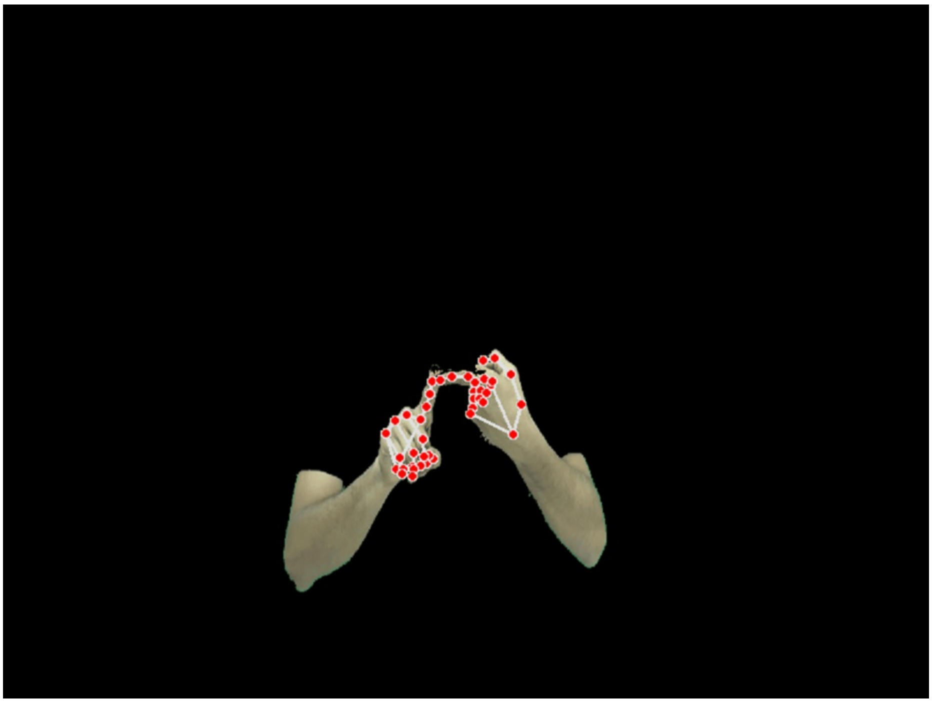


FIGURE 4: Hand landmark detection using Google MediaPipe holistic model.

Gesture Animation

The hand parameters (coordinates of the landmarks) detected in all the key frames of the input gesture video form the complete description of the corresponding sign and are stored in the database. For a query sign, the corresponding gesture video is created by animation using this stored hand parameter information rather than playing back a stored real gesture video. Thus, our proposed framework helps in doing away with large memory requirement.

In our method for animating the gestures, we propose to use interpolation using image metamorphosis to synthesize the in-between frames between two key frames [38]. This is done by applying image warping followed with the intensity interpolation. Intensity interpolation helps in assigning the intensity values to the spatially transformed pixels. For image warping, we need to find a transformation function to generate the intermediate key points between the feature points of two consecutive key frames or images. A generalized expression for transforming the geometric co-ordinates can be explained as follows:

$$(\mathbf{X_{transformed}}, \mathbf{Y_{transformed}}) = T(\mathbf{X_{original}}, \mathbf{Y_{original}})$$

where T is the transformation function used to transform the pixel co-ordinates ($\mathbf{X_{original}}$, $\mathbf{Y_{original}}$) of the original image to the corresponding pixel co-ordinates ($\mathbf{X_{transformed}}$, $\mathbf{Y_{transformed}}$) in the transformed image.

In the extracted key frames, we earlier obtained the hand key points as our feature points using Google MediaPipe. These key points are in the form of geometric co-ordinates. In our work, to find the intermediate key points between the key points or feature points of two key frames, we use spline interpolation using not-a-knot end condition. These intermediate key points are the interpolated values obtained using spline interpolation. The process of image morphing gets completed by assigning the intensity values to the co-ordinates of the intermediate key points. We do this with the help of intensity interpolation. For intensity interpolation, we have used bicubic interpolation. In bicubic interpolation, the intensity value assigned to the point (X,Y) is given by the following equation:

$$\mathbf{I(X,Y)} = \sum_{\mathbf{j=0}}^{\mathbf{3}} \sum_{\mathbf{k=0}}^{\mathbf{3}} \mathbf{c_{jk} X^j Y^k}$$

where the co-efficients $\mathbf{c_{jk}}$ are determined using the equations written using the 16 nearest neighbors of the point **(X,Y)**. This is how the in-between frames for two key frames are synthesized. Our proposed approach resulted in producing smooth and natural gesture animation videos.

Standalone Application Development

After successfully performing a smooth and natural gesture animation framework, we developed a standalone MATLAB application (app) for the purpose of educating the community with hearing and speech impairments. We developed this application using MATLAB App Designer but can be successfully installed on any system without the requirement of MATLAB on that system. We developed the system for generating animated sign gestures of ISL fingerspelling (that included digits and alphabets of ISL) and some basic words in ISL. A screenshot of the application displayed on the desktop monitor is shown in Figure 5. In this app, corresponding to the user input text, an animated gesture video will be played on the screen, as depicted in Figure 6. The graphical user interface of the app has the provision for the user to play the animation in slow speed or fast speed. Also, there is a provision to pause and resume the application whenever desired.

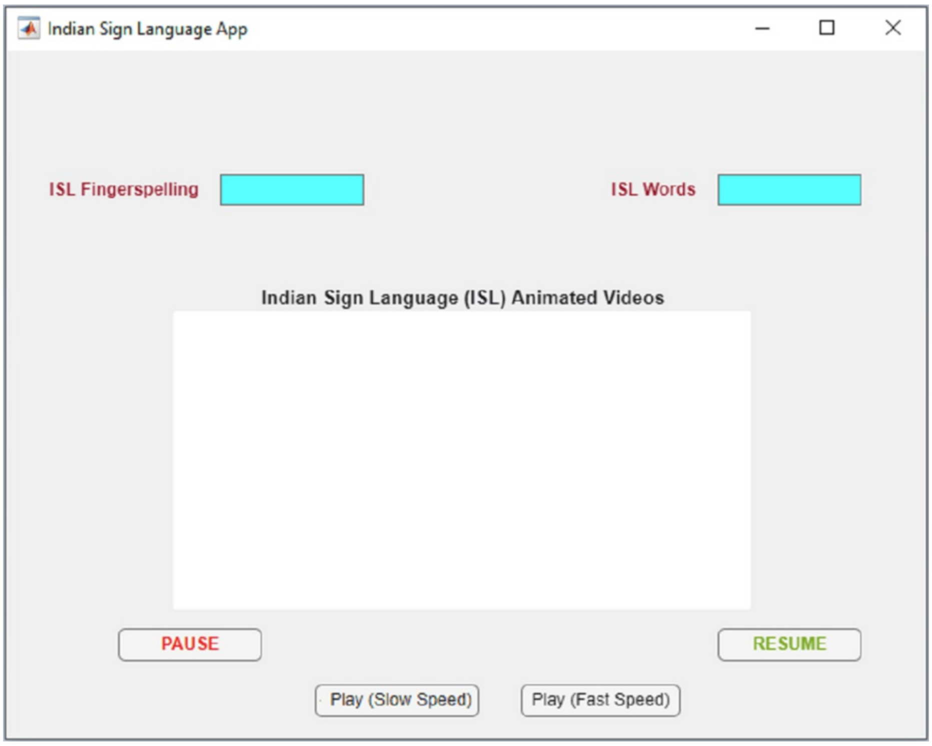
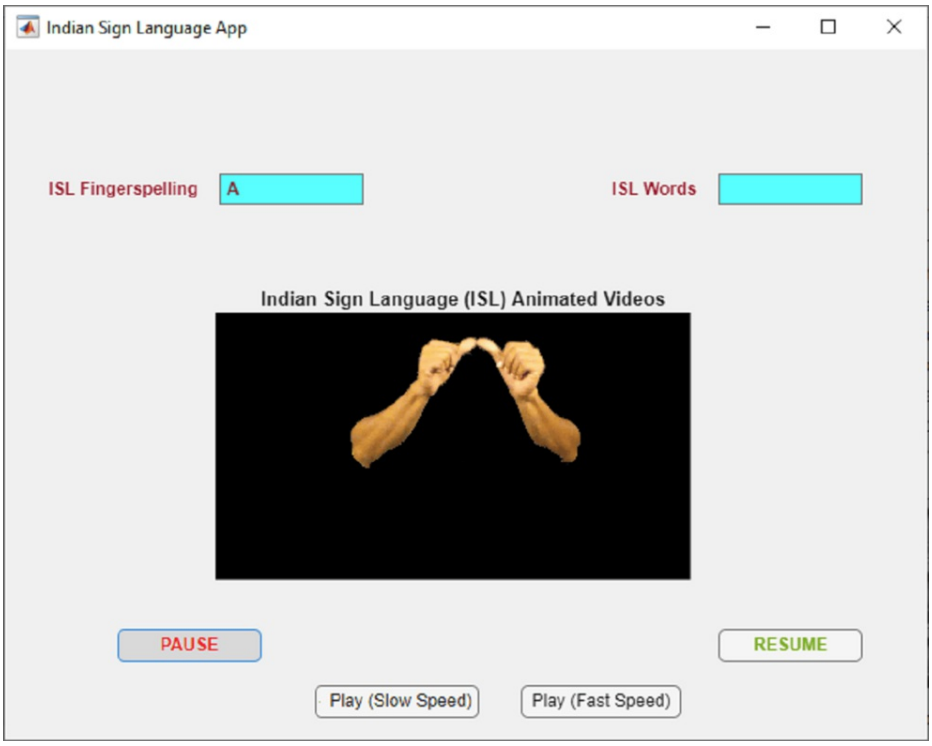


FIGURE 5: Screenshot of the developed application.



Indian Sign Language App

ISL Fingerspelling: A

ISL Words:

Indian Sign Language (ISL) Animated Videos

FIGURE 6: One instance of the synthesized gesture video while performing animation of ISL alphabet A.

Results

We performed our experiments on a system with Intel(R) Xeon(R) W-1290P CPU @ 3.70 GHz processor with 32.00 GB RAM and an NVIDIA Quadro P2200 GPU of 5 GB RAM. We used MATLAB R2022b to carry out hand region extraction, key frame extraction, gesture animation and application development. We used OpenCV and Python for extracting hand parameters from the key frames. We performed our experiments on both static and dynamic gestures of ISL fingerspelling and ISL words. Results of hand region extraction, key frame extraction and hand parameter extraction are shown in Figures *2*-*4*. These figures demonstrate the efficiency of our proposed method.

In Figure *7*, we show the results of our proposed gesture animation method. This figure shows some key frames and the generated in-between frames. Figures *7*(a), (f), (k) and (p) are the synthesized key frames of the ISL alphabet W. Figures *7*(b)-(e) are the synthesized in-between frames between key frames Figures *7*(a) and (f). Figures *7*(g)-(j) depict the generated frames between the key frames Figures *7*(f) and (k). Similarly, for another key frame pair shown in Figures *7*(k) and (p), Figures *7*(l)-(o) are the synthesized in-between frames. As illustrated in this figure, our proposed method is efficient and produces smooth and natural gesture.

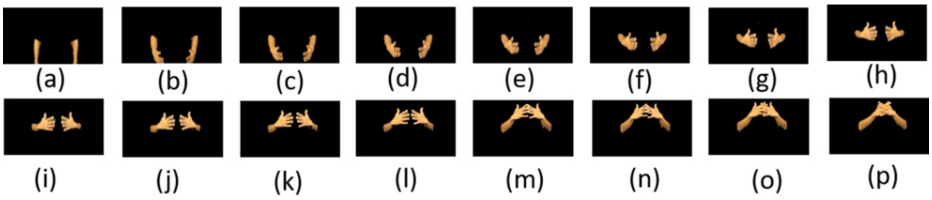


FIGURE 7: (a), (f), (k) and (p) are four key frames of ISL alphabet W (also depicted in Figure 3), whereas (b)-(e), (g)-(j) and (l)-(o) are generated animated frames between the key frame pairs (a) and (f), (f) and (k), and (k) and (p), respectively.

The above example shown in Figure *7* is of a static sign gesture in which the hands move from rest position, make the sign and then return to the rest position. Figure 8 shows results of our proposed gesture animation method for a dynamic gesture, which also illustrates the efficiency of our proposed approach in terms of smoothness and naturalness of the synthesized gesture video.

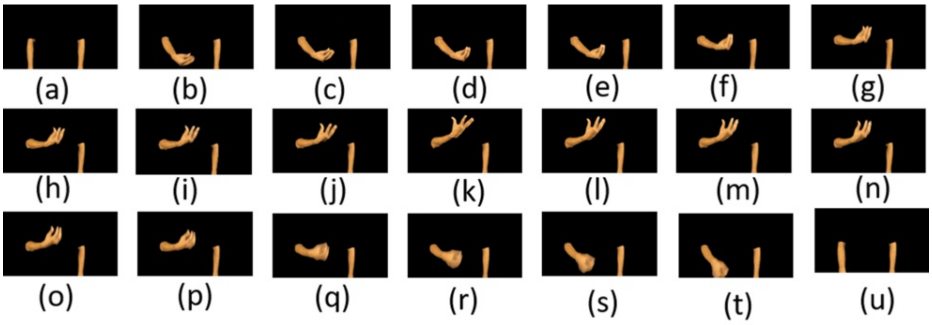


FIGURE 8: (a), (f), (k), (p) and (u) are five key frames of an ISL word 'Morning', whereas (b)-(e), (g)-(j), (l)-(o) and (q)-(t) are synthesized animated frames between the key frame pairs (a) and (f), (f) and (k), (k) and (p), and (p) and (u), respectively.

Discussion

In this work, we have developed an animation-based e-learning app to facilitate education of ISL. Major steps of our method included hand region extraction, key frame extraction, hand parameter estimation, gesture animation and app development. In this work, all of these steps have been efficiently performed. Our proposed method of hand region extraction performed well and extracted hand regions clearly from the image frame. K-means clustering-based unsupervised classification and segmentation helped us in separating out the hands efficiently. The two metrics used for key frame extraction complemented each other to summarize the entire gesture video into fewer key frames compared to number of key frames provided by them individually. The threshold values for both the metrics are empirically calculated. Google MediaPipe being an efficient tool for hand parameters extraction provided the parameter values without any difficulty. It worked well for single-handed as well as double-handed signs. With the help of spline interpolation and intensity interpolation, smooth and efficient gesture synthesis and animation part is performed. Spline interpolation helped in achieving the smoothness of the synthesis part, whereas the intensity interpolation helped in maintaining the color part of the synthesized frames. The standalone application development using MATLAB App Designer helped in developing the e-learning tool for ISL. This tool plays ISL fingerspelling and some ISL words. This app is prepared to provide education of ISL to the hearing- and speech-impaired community in India. However, this can be used by anyone to learn ISL. This will lead to the inclusive growth of society.

In Table 2, based on the mean opinion score, we present the evaluation results of the animated sign language videos. A total of 15 users took part in this evaluation process. Those users were first shown the original sign videos and then animated videos. Next, based on three criteria, namely, smoothness of the animated videos, similarity with the original sign videos and clarity in understanding, evaluation scores are obtained. The scores are obtained on a scale of 0 to 5 (0 being the lowest and 5 being the highest). We obtained scores on ISL alphabets and digits, ISL fingerspelling and ISL words separately. From Table 2, we observe that generated sign language videos based on our proposed approach performed well. We also collected feedback from the users regarding various aspects of our developed application.

Categories	Mean Opinion Score		
	Smoothness	Similarity	Clarity in Understanding
ISL Alphabets and Digits	4.2	4.7	4.0
ISL Fingerspelling	4.2	4.6	3.9
ISL Words	3.9	4.7	3.9

TABLE 2: Evaluation Results of the Animated ISL Videos Based on Mean Opinion Score

Conclusions

In this paper, an e-learning tool is developed for online teaching of ISL for the community in India suffering from hearing and speech impairments. A standalone MATLAB-based app for the same is developed in which sign gestures in ISL are synthesized by animation. For a query sign, the corresponding gesture is animated by creating a sequence of video frames containing 3D hand shapes with desired palm orientation and finger joint angles as per the hand parameter information extracted from a real gesture video. In our proposed method, rather than extracting hand parameter information from every frame in the input real video and synthesizing the same, we do this only for a few number of key frames in the video. Frames between the key frames are generated by interpolation using image metamorphosis. Thus, our proposed method creates smooth and natural-looking animated videos from a comparatively less amount of stored gesture information, thereby offering large savings in memory compared to most other existing online sign language dictionaries and sign language learning softwares that just play back stored real gesture videos. Our application has been developed for ISL fingerspelling that includes digits, alphabets and combination of these, and also for some ISL words of daily usages. The app works equally well for static and dynamic gestures. To test the efficacy of our proposed approach, we performed user-based evaluation. Based on the mean opinion score obtained from the users, we found our proposed method and developed application to be efficient and effective.

To the best of our knowledge, this is the first animation-based e-learning tool developed for teaching and learning of ISL. In India, where special schools are located only in some cities and big towns, this app for computer-aided sign language teaching and learning will be a boon to the community with hearing and speech impairments for learning sign language, even while seated at home. The same can also be used by teachers at special schools or integrated schools for teaching sign language as well as language skill development among students with hearing and speech impairments. Further work on this may include text-to-sign language translation that can be used for bilingual dictionaries, signing books and for computer-generated translations of printed text into a sign language. The software may also be extended to speech-to-sign language translation and be used by the people with hearing and speech impairments for understanding what others speak while conversing with normal beings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Navneet Nayan

Acquisition, analysis, or interpretation of data: Navneet Nayan, Debashis Ghosh, Pyari M. Pradhan

Drafting of the manuscript: Navneet Nayan

Critical review of the manuscript for important intellectual content: Debashis Ghosh, Pyari M. Pradhan

Supervision: Debashis Ghosh, Pyari M. Pradhan

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.
Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.
Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** Department of Science & Technology, Government of India, Grant No. SEED-TIDE-055-2016. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

Our sincere thanks go to the Department of Science and Technology, Government of India, for providing financial support in carrying out this project on developing this app for e-learning of ISL to be used by the hearing-impaired children in India. This work could not have been possible without the ISL dictionary videos that have been sourced from the Indian Sign Language Dictionary 1st edition developed by the Indian Sign Language Research and Training Centre (ISLRTC), Department of Empowerment of Persons with Disabilities (DEPwD), Ministry of Social Justice and Empowerment, Government of India. We would also like to acknowledge Kumar Shubham Singh and Athar Ali for helping in preparing the ISL fingerspelling and ISL words dataset for this work.

References

1. Buttussi F, Chittaro L, Coppo M: Using Web3D technologies for visualization and search of signs in an international sign language dictionary. Proceedings of the Twelfth International Conference on 3D Web Technology. 2007, 61-70. 10.1145/1229390.1229401
2. Cavender AC, Otero DS, Bigham JP, Ladner RE: ASL-STEM forum: Enabling sign language to grow through online collaboration. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010, 2075-2078.
3. Han J, Awad G, Sutherland A: Boosted subunits: A framework for recognizing sign language from videos. IET Image Processing. 2013, 7:70-80. 10.1049/iet-ipr.2012.0273
4. Dasgupta T, Shukla S, Diwakar S, Basu A: A multilingual multimedia Indian sign language dictionary. Proceedings of the 6th Workshop on Asian Language Resources. 2008, 57-64.
5. Gebre BG, Wittenburg P, Heskes T: Automatic sign language identification. 2013 IEEE International Conference on Image Processing. 2013, 2626-2630. 10.1109/ICIP.2013.6738541
6. Geitz S, Hanson T, Maher S: Computer generated 3-dimensional models of manual alphabet handshapes for the World Wide Web. Proceedings of the Second Annual ACM Conference on Assistive Technologies. 1996, 27-31. 10.1145/228347.228353
7. Patel I, Rao YS: Technologies automated speech recognition approach to finger spelling. 2010 Second International Conference on Computing, Communication and Networking Technologies. 2010, 1-6. 10.1109/ICCCNT.2010.5591724
8. Seki Y, Sato T: A training system of orientation and mobility for blind people using acoustic virtual reality. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2011, 19:95-104. 10.1109/tnsre.2010.2064791
9. Moustakas K, Tzovaras D, Dybkjaer L, Bernsen N, Aran O: Using modality replacement to facilitate communication between visually and hearing-impaired people. IEEE MultiMedia. 2010, 18:26-37. 10.1109/MMUL.2010.22
10. Sedgwick E, Alkoby K, Davidson MJ, et al.: Toward the effective animation of American Sign Language. Procedings of the 9th International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media. 2001, 375-378.
11. Kennaway R: Synthetic animation of deaf signing gestures. Gesture and Sign Language in Human-Computer Interaction. Wachsmuth I, Sowa T (ed): Springer, Berlin, Heidelberg; 2002. 2298:146-157. 10.1007/3-540-47873-6_15
12. Verlinden M, Zwitserlood I, Frowein H: Multimedia with animated sign language for deaf learners. Proceedings of Ed-Media. 2005, 4759-4764.
13. Söylev A, Mendi E: Turkish Sign Language animation with motion capture. 2014 22nd Signal Processing and Communications Applications Conference. 2014, 834-837. 10.1109/SIU.2014.6830359
14. Kopp S, Wachsmuth I: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds. 2004, 15:39-52. 10.1002/cav.6
15. Olivier P: Gesture synthesis in a real-world ECA. Affective Dialogue Systems. André E, Dybkjær L, Minker W, Heisterkamp P (ed): Springer, Berlin, Heidelberg; 2004. 3068:319-322. 10.1007/978-3-540-24842-2_35
16. Hartmann B, Mancini M, Pelachaud C: Implementing expressive gesture synthesis for embodied conversational agents. Gesture in Human-Computer Interaction and Simulation. Gibet S, Courty N, Kamp JF (ed): Springer, Berlin, Heidelberg; 2006. 188-199. 10.1007/11678816_22
17. Huenerfauth M, Zhou L, Gu E, Allbeck J: Design and evaluation of an American Sign Language generator. Proceedings of the Workshop on Embodied Language Processing. 2005, 51-58. 10.3115/1610065.1610072
18. Wik P, Hjalmarsson A: Embodied conversational agents in computer assisted language learning. Speech Communication. 2009, 51:1024-1037. 10.1016/j.specom.2009.05.006
19. Sáfár É, Marshall I: The architecture of an English-text-to-Sign-Languages translation system. Recent Advances in Natural Language Processing. 2001, 223-228.
20. San Segundo R, Montero JM, Macías-Guarasa J, Córdoba R, Ferreiros J, Pardo JM: Generating gestures from speech. Proceedings of the 8th International Conference on Spoken Language Processing. 2004, 1817-1820.
21. Sarkar B, Datta K, Datta CD, et al.: A translator for Bangla text to sign language. 2009 Annual IEEE India Conference. 2009, 1-4. 10.1109/INDCON.2009.5409449
22. Almohimeed A, Wald M, Damper R: Arabic text to Arabic Sign Language translation system for the deaf and hearing-impaired community. Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies. 2011, 101-109.
23. San-Segundo R, Montero JM, Córdoba R, et al.: Design, development and field evaluation of a Spanish into sign language translation system. Pattern Analysis and Applications. 2012, 15:203-224.10.1007/s10044-011-0243-9
24. Lopez-Ludena V, San-Segundo R, Martin R, Sanchez D, Garcia A: Evaluating a speech communication system for deaf people. IEEE Latin America Transactions. 2011, 9:565-570. 10.1109/TLA.2011.5993744
25. Verma V, Ghosh D: Hand gesture reconstruction and animation. Proceedings of the 2nd Indian International Conference on Artificial Intelligence. 2005, 537-555.
26. Verma V, Ghosh D: Synthesis and animation of dynamic hand gestures for sign language generation. Journal of Intelligent Systems. 2008, 17:173-184. 10.1515/JISYS.2008.17.1-3.173
27. Shankar VN, Ghosh D: Dynamic hand gesture synthesis and animation using image morphing technique. 2006 IET International Conference on Visual Information Engineering. 2006, 543-548.
28. Bhuyan MK, Narra C, Chandra DS: Hand gesture animation by key frame extraction. 2011 International Conference on Image Information Processing. 2011, 1-6. 10.1109/ICIIP.2011.6108947
29. Bhuyan MK, Ramaraju VV, Iwahori Y: Hand gesture recognition and animation for local hand motions. International Journal of Machine Learning and Cybernetics. 2013, 5:607-623. 10.1007/s13042-013-0158-4
30. Kaur S, Singh M: Indian Sign Language animation generation system. 2015 1st International Conference on Next Generation Computing Technologies (NGCT). 2015, 909-914. 10.1109/NGCT.2015.7375251

31. Wiles O, Koepke AS, Zisserman A: X2face: A network for controlling face generation using images, audio, and pose codes. Computer Vision - ECCV 2018. Ferrari V, Hebert M, Sminchisescu C, Weiss Y (ed): Springer, Cham; 2018. 690-706. 10.1007/978-3-030-01261-8_41
32. Liu Y, De Nadai M, Zen G, Sebe N, Lepri B: Gesture-to-gesture translation in the wild via category-independent conditional maps. Proceedings of the 27th ACM International Conference on Multimedia. 2019, 1916-1924. 10.1145/3343031.3351020
33. Toledano O, Marmor Y, Gertz D: Image animation with keypoint mask. arXiv Preprint. 2021, arXiv:2112:10457.
34. Tao J, Wang B, Ge T, Jiang Y, Li W, Duan L: Motion transformer for unsupervised image animation. Computer Vision - ECCV 2022. Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (ed): Springer, Cham; 2022. 13676:702-719. 10.1007/978-3-031-19787-1_40
35. Simon T, Joo H, Matthews I, Sheikh Y: Hand keypoint detection in single images using multiview bootstrapping. 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017, 4645-4653. 10.1109/CVPR.2017.494
36. Lugaresi C, Tang J, Nash H, et al.: MediaPipe: A framework for perceiving and processing reality. Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR). 2019.
37. Viola P, Jones M: Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2001, 511-518. 10.1109/CVPR.2001.990517
38. Wolberg G: Image morphing: A survey. The Visual Computer. 1998, 14:360-372. 10.1007/s003710050148