Article

Not peer-reviewed version

Signsability: Enhancing Communication Through Sign Language App

Din Ezra *, Shai Mastitz *, Irina Rabaev *

Posted Date: 2 August 2024

doi: 10.20944/preprints202408.0125.v1

Keywords: Sign Language Recognition; Deep Learning; Computer Vision; MediaPipe; Video Processing; ISL Dataset; Web App

Article

Signsability: Enhancing Communication Through Sign Language App

Din Ezra *,‡, Shai Mastitz *,‡ and Irina Rabaev

Software Engineering Department, Shamoon College of Engineering, 56 Bialik St.,Be'er Sheva 8410802, Israel; ezradi@ac.sce.ac.il (D.E.); shaima7@ac.sce.ac.il (S.M.); irinar@ac.sce.ac.il (I.R.)

* Correspondence: dinezra11@gmail.com (D.E.); shaimastitz1@gmail.com (S.M.)

‡ These authors contributed equally to this work.

Abstract: The integration of sign language recognition systems into digital platforms has the potential to bridge communication gaps between the deaf community and the broader population. This paper introduces an advanced Israeli Sign Language (ISL) recognition system designed to interpret dynamic motion gestures, addressing a critical need for more sophisticated and fluid communication tools. Unlike conventional systems that focus solely on static signs, our approach incorporates both Deep Learning and Computer Vision techniques to analyze and translate dynamic gestures captured in real-time video. We provide a comprehensive account of our preprocessing pipeline, detailing every stage from video collection to the extraction of landmarks using MediaPipe, including the mathematical equations used for preprocessing these landmarks and the final recognition process. The dataset utilized for training our model is unique in its comprehensiveness and is publicly accessible, enhancing the reproducibility and expansion of future research. The deployment of our model on a publicly accessible website allows users to engage with ISL interactively, facilitating both learning and practice. We discuss the development process, the challenges overcome, and the anticipated societal impact of our system in promoting greater inclusivity and understanding.

Keywords: sign language recognition; deep learning; Computer Vision; MediaPipe; video processing; ISL Dataset; Web App

1. Introduction

When visiting a foreign country for either vacation or business, one often experiences significant challenges in communicating with the local environment. Simple tasks can become exceedingly complicated when one is unable to express her or his ideas or emotions effectively.

Today, this complication has become much more subtle due to the major technological progress we encounter in the development of translation software and language learning platforms. These developments have diminished the necessity of mastering foreign languages or relying on translators. Consequently, individuals worldwide can effortlessly establish meaningful connections despite speaking entirely different languages. However, there is still a communication gap that remains unresolved: the barrier between hearing individuals and the deaf community.

Sign language is a complex form of visual communication utilized primarily by individuals who are deaf or hard of hearing. It encompasses hand shapes, facial expressions, and body movements to convey messages. Facial expressions are particularly significant, as they add nuances and convey emotions, enriching the overall communication. Sign language interpreters facilitate interactions between deaf and hearing individuals, serving as vital mediators. Sign languages are integral to the identity and culture of deaf communities, providing essential accessibility in education, information dissemination, and daily interactions. Therefore, proficiency in sign language or the ability to translate signs into spoken language is crucial for facilitating interactions between hearing and hard of hearing individuals.

For each natural spoken language, there exists a distinct sign language counterpart. For instance, American Sign Language (ASL) corresponds to spoken English in the United States, while Israeli Sign Language (ISL, also called 'Shassi') corresponds to Hebrew spoken in Israel. According to K. Cheng [1],

 doi:10.20944/preprints202408.0125.v1

ASL and ISL differ significantly from one another in both the meanings of signs and their grammatical structures.

The distinctions among various sign languages necessitate tailored technological solutions specific to each language. Solutions designed for Chinese sign language, for instance, cannot be seamlessly applied to Spanish sign language. Therefore, the development of a system intended to aid or assist the deaf and hard of hearing population in Israel must be custom-tailored to meet the unique linguistic and cultural requirements of the country.

This study focuses on developing and deploying a deep-learning model for the classification and recognition of sign language, specifically focusing on Israeli Sign Language. 'Signsability' is a unique system developed by our team designed to enhance the communication between hearing and hard of hearing individuals by offering two primary functionalities:

1. Translation: The system efficiently translates videos of Israeli Sign Language signs and gestures into written Hebrew.
2. Learning: The system allows users to practice ISL and receive immediate feedback on the accuracy of their signed words. Utilizing a web camera, the application captures the signs, processes them through a trained classifier, and provides users with information on the correctness of their signing.

To conduct this study, we first compiled a dataset of videos featuring words in ISL form with corresponding Hebrew translations. The significance of this dataset extends beyond our immediate research, offering a valuable resource for further developments in sign language translation technologies. This dataset is publicly available on Kaggle, ensuring that other researchers can benefit from and build upon our work. Second, we trained an end-to-end deep-learning model on the compiled dataset. Third, we deployed the trained model into a real-time web-based platform-independent application.

Given the limited research on ISL, we believe this study makes significant contributions to both the hearing and deaf communities. By developing a model that accurately classifies and transcribes videos of signers into written Hebrew, we aim to bridge the communication gap between these communities, facilitating more effective and inclusive interactions. We believe that both the dataset and the real-time application will serve as a foundation for future research in this field. [1]

2. Related Work

This section overviews key approaches and methodologies of the current state of research in the field of automatic sign language classification and recognition.

Murali et al. [2] developed a classifier to recognize letters in American Sign Language. The model was specifically trained to identify ten letters: a, b, c, d, h, k, n, o, t, and y. The training dataset comprised 2,000 images collected from two individual ASL speakers. The model demonstrated high accuracy, achieving over 90% for each of the trained letters. Obi et al. [3] presented an experiment utilizing an approach similar to [2] but addressing the problem on a broader scale. Their study aimed to effectively classify all 26 letters of the English alphabet. The goal was to develop a system capable of recognizing individual letters sequentially to spell out entire words. Unlike recognizing whole words as gestures, this system focused on recognizing and interpreting the spelling of words. The authors reported a 96.3% accuracy in classifying the letters. Despite this high accuracy, the authors identified a significant limitation of the model. For the model to perform as expected and achieve the desired accuracy, each letter must be displayed for a relatively extended period, one after the other. This requirement implies that the model cannot recognize a sequence of letters signed at the natural rhythm of real speech, rendering the solution impractical for real-world applications. In their research, Balaha et al. [4] focused on Arabic Sign Language and described the integration of Convolutional Neural

[1] The study was conducted as a part of the undergraduate project for a B.Sc. degree in Software Engineering.

 doi:10.20944/preprints202408.0125.v1

Network (CNN) and Recurrent Neural Network (RNN) architectures to train a classification model for 20 words in the Arabic language. This study employed RNNs to recognize words articulated through continuous motion in sign language rather than static signs. To train the model, approximately 8,467 videos of the 20 words were recorded by 72 volunteers. The accuracy for classifying the words ranged from 72%-100%. Gogoi [5] proposed a slightly different approach to solving the problem. Instead of training the classifier directly on image data, they first extracted the landmarks of body parts and gestures from the videos, subsequently feeding these coordinates into the deep-learning model. For landmark extraction, the authors utilized the MediaPipe library, an open-source library developed by Google that includes various pre-trained models for time-series data processing tasks, such as video and audio processing. The authors focused on nine letters in Assamesse and achieved 99% accuracy on the test dataset.

The primary limitations identified in the studies reviewed above relate to the use of static and isolated gestures. This is a critical issue, as dynamic movements in sign language are essential for conveying subtle interpretations. Shamitha and Badarinath [6] presented an alternative approach that more effectively addresses this issue. Their model also utilizes the MediaPipe library for landmark extraction; however, the key difference lies in their use of a Recurrent Neural Network architecture. This architecture facilitates the classification of sequential data, enabling the model to interpret the motion and movement inherent in sign language gestures. Consequently, the model can recognize dynamic patterns rather than being confined to static signs.

3. Signsability Dataset

As with any comprehensive data science project, our initial phase involved data collection. To conduct the study, we compiled and annotated the 'Signsability' dataset. Given that none of the developers possess fluency in Israeli Sign Language, our objective was to collaborate with ISL speakers. Consequently, it was imperative to design the data collection process to be highly accessible to individuals without a background in software engineering. Therefore, we developed the application "Signsability: Data Collector" to facilitate this collaboration.[2]

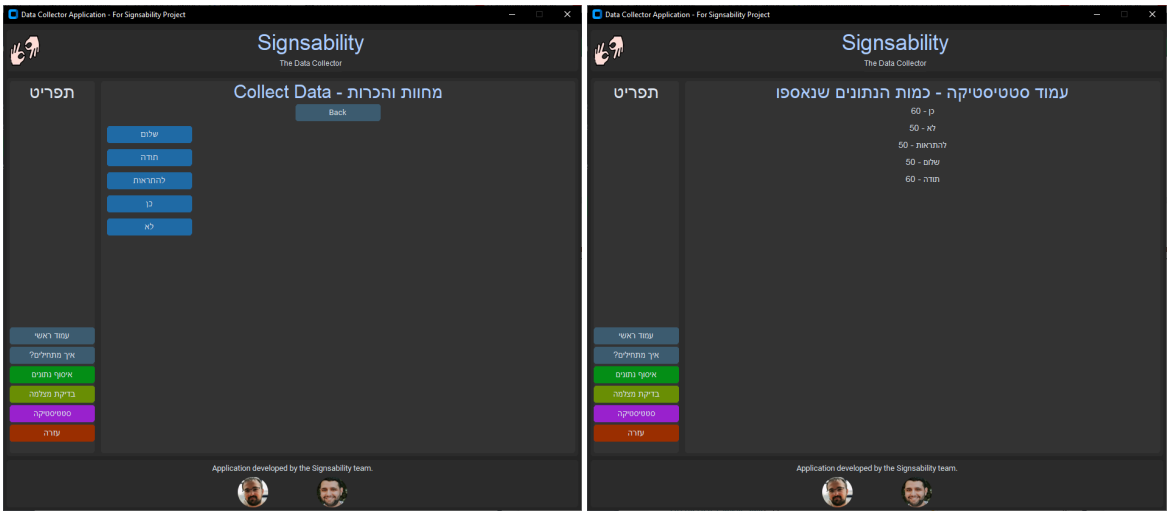
The "Signsability: Data Collector" application features an intuitive and user-friendly graphical interface. As illustrated in Figure 1 (a), the application allows the data contributor to press a button corresponding to the specific word for which data is being collected. The application then begins recording the signs performed by the contributor via the webcam. Subsequently, the application executes the predefined preprocessing pipeline and stores the data in a suitable format that can directly be fed into the deep-learning model to be trained. The suitable format and preprocessing pipeline are described in detail in Section 4.2.

Additionally, a statistical feature has been incorporated into the application, displaying the quantity of data collected for each individual sign. This functionality facilitates a balanced and consistent data collection process, preventing discrepancies in data size per label and mitigating the risk of biased data. The statistical window of the application is shown in Figure 1 (b).

---

[2] The application will be provided at no cost to any individual who requests it via email. We will gladly fulfill all such requests.

 doi:10.20944/preprints202408.0125.v1



(a) **(b)**

Figure 1. Screenshots from data collection application developed by our team. (a) Gallery of words and letters to be collected; (b) a screenshot of the statistical window.

At the time of writing this paper, the 'Signsability' dataset comprises a total of 600 images and 400 videos. The 600 images are categorized into three different letters — א, ב, and ג, with 200 samples each. The signs for these letters are static, thus images are sufficient to capture and recognize them. The videos, on the other hand, are for words that involve motion and dynamic gestures in their sign language pronunciation. The 400 collected videos capture five distinct words, with 80 samples each. Each video was recorded using a simple web camera. The average duration of the sample videos is 3 seconds, which are then split into 30 frames.

For the purposes of training, the dataset was split using an 80-20-20 ratio for the training, validation, and testing sets, respectively.

The collected data has been published in the Kaggle datasets section following initial processing with MediaPipe.[3] Aside from landmark extraction, the data remains unnormalized and unprocessed to allow other researchers to explore, process, and manipulate it as they see fit. The data collection process is still ongoing, and the dataset is regularly updated with the objective of creating a comprehensive, publicly available resource.

4. Methodology

The method consists of three primary phases. In the first phase, given a video or a static image of the sign as input, the landmarks corresponding to various body parts are extracted and converted into coordinate points. The second phase involves preprocessing the landmark coordinates, which includes converting their position relative to a reference anchor landmark, calculating the distances among them, and applying normalization techniques. In the final phase, the processed landmarks are fed into a deep-learning classifier.

4.1. Landmarks Extraction

Body posture recognition and the extraction of limbs, such as hands and fingers, from images play a critical role in understanding human movements and gestures. For the task of limbs' extraction, we utilized MediaPipe[4] – an open-source library developed by Google. By leveraging MediaPipe's

[3] The 'Signsability' dataset is available for download on Kaggle: https://kaggle.com/datasets/fe1ca51ef2c13347756617aedfc611d5699035cdc74ca72e89a357e466f8ebb5

[4] https://ai.google.dev/edge/mediapipe/solutions/guide

 doi:10.20944/preprints202408.0125.v1

pre-trained models, we can accurately detect and extract keypoints from various body parts, facilitating precise body posture recognition. These keypoints, known as landmarks, are crucial for interpreting the positions and movements of limbs in an image. In the context of sign language recognition, for instance, accurately extracting the positions of hands and fingers is essential for correctly interpreting gestures. MediaPipe's efficient processing capabilities enable real-time analysis and extraction of these keypoints, providing a solid foundation for subsequent classification tasks. This initial step of recognizing and extracting body posture elements using MediaPipe ensures that the following stages of gesture interpretation and word classification are based on accurate and reliable data.

4.2. Preprocessing

The landmarks extracted in the first phase are represented by (X, Y) coordinates. An image consists of multiple landmarks that represent the body gesture. After the initial extraction, we perform a series of preprocessing operations. These operations include converting landmarks' positions relative to a reference anchor landmark, calculating the distances between them, and applying normalization techniques (see Equations 1 - 4).

$$Distances_i(x, y) = \begin{cases} X_i = X_i - X_0 \\ Y_i = Y_i - Y_0 \end{cases} \tag{1}$$

$$L = \{X_i, Y_j : X_i, Y_j \in Distances\} \tag{2}$$

$$\alpha = max(l : l \in L) \tag{3}$$

$$Landmark_i(x, y) = \begin{cases} X_i = Distances_i(X)/\alpha \\ Y_i = Distances_i(Y)/\alpha \end{cases} \tag{4}$$

In the initial stage of the preprocessing pipeline, we calculate the distance between the coordinates of each landmark, represented as X and Y points, and the coordinates of the first landmark, such as the thumb (Equation 1). After that, we concatenate the X and Y coordinates of all calculated distances into a comprehensive list and identify the maximum value, as demonstrated in Equations 2 and 3. Finally, each distance is divided by the maximum value, thereby normalizing the entire dataset and preparing it for the model's training process (Equation 4).

4.3. Training

After the dataset was preprocessed and prepared, we proceeded with the training session for the Israeli Sign Language recognition model. Throughout our research, we developed and trained two different models: one for classifying a sign to its corresponding Hebrew word based on a static image, and another for recognizing an entire gesture and mapping it to its respective sign language term from a video. We describe these models in Sections 4.3.1 and 4.3.2, respectively.

4.3.1. Static Signs Recognition - MediaPipe with a Simple NN

The initial model we trained was a basic neural network composed solely of dense and dropout layers. The training dataset comprised a total of 600 images representing the sign patterns for three Hebrew letters: (a) א, (b) ב, and (c) ג, with 200 images per class. By leveraging the robust capabilities of MediaPipe, we were able to utilize the numerical representation of the extracted landmarks to construct a relatively simple yet effective neural network. We opted to develop the model from scratch, obviating the need for reliance on pretrained models. The network architecture comprises three dense layers with ReLU activation functions, having two dropout layers between them. The model contains a total of 1,114 trainable parameters. For the loss function, we selected sparse categorical cross-entropy, enabling the model to classify a single word for each sample.

The classification pipeline for the static sign recognition is shown in Figure 2.

 doi:10.20944/preprints202408.0125.v1

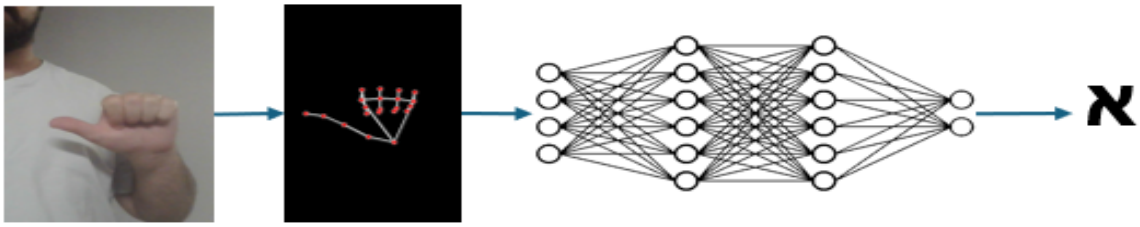


Figure 2. Classification pipeline with MediaPipe for static signs.

4.3.2. Dynamic Signs Recognition - MediaPipe with RNN

Dynamic signs recognition involves the identification of hand movements and gestures over a sequence of images. For this task, we utilized an RNN model, which is designed to handle sequential data, making it ideal for capturing the temporal dynamics of hand movements.

The entire classification pipeline for dynamic gesture recognition from video is illustrated in Figure 3. Initially, frames are extracted from the video, with each video comprising a total of 30 frames. Subsequently, each frame is processed using the MediaPipe model to extract landmarks specific to that frame. These extracted landmarks are then compiled into a sequence representing the entire video. This generated sequence is subsequently fed into the RNN model with Long Short-Term Memory (LSTM) layers, which classify the sequence of frames into a corresponding word.

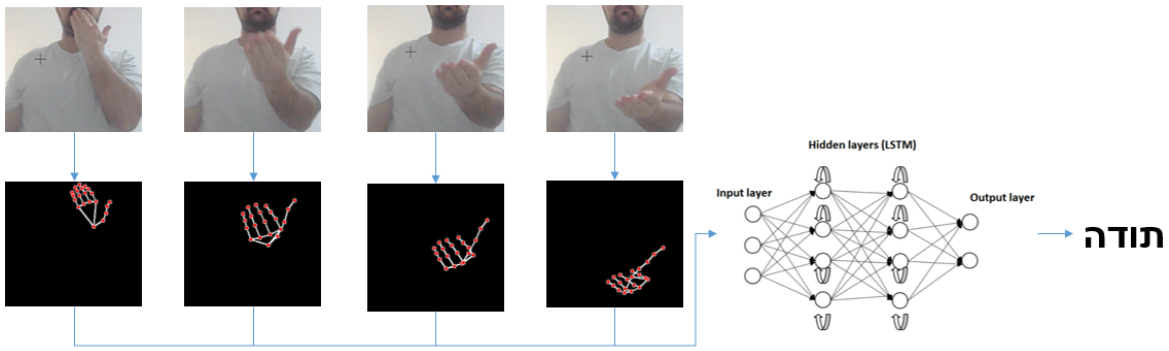


Figure 3. Classification pipeline with MediaPipe and RNN.

The input samples provided to the neural network are structured with dimensions (1, 30, 1663), where each sample represents a video segment containing 30 frames, and each frame comprises 1663 preprocessed landmarks. The architecture of the trained model consists of three LSTM layers, followed by three Dense layers. The total number of trainable parameters in this version of the model is 596,741. The architecture summary is shown in Figure 4.

 doi:10.20944/preprints202408.0125.v1

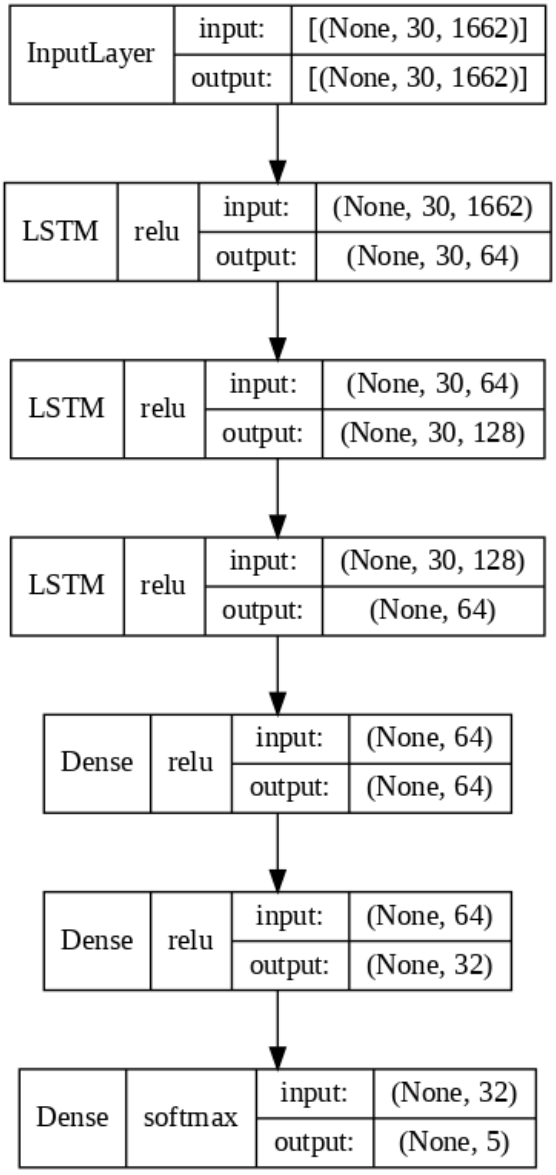


Figure 4. Illustration of the RNN model architecture.

5. Results and Discussions

In our project, as is customary when working with data, particularly new data that is being collected concurrently with the development of the classification network, we decided to evaluate several approaches to select the optimal classification network. Our goal is to achieve results that align with the project's requirements, defined during the project's requirement specification phase. We examined two different types of classification networks: Neural Networks for classifying a sign from a static image and Recurrent Neural Networks for dynamic sign classification from video. After obtaining the results, we selected the optimal network based on accuracy and the need for a network capable of handling diverse data, including images and videos.

5.1. Static Signs Recognition Results

The obtained classification accuracy for static sing recognition was 0.96. The precision, recall, and F-1-score for each class are listed in Table 1.

 doi:10.20944/preprints202408.0125.v1

Table 1. Classification report for the test dataset, with the letters א, ב, ג.

Class	P	R	F1-score
א	1	0.99	099
ב	0.98	0.92	0.95
ג	0.91	0.99	0.95
Average	0.96	0.96	0.96

As can be seen, the model achieved high accuracy across all classes. However, it is important to note that all classes represent static hand patterns in sign language, devoid of motion or dynamic gestures. As previously mentioned, the ability to recognize motion and dynamic gestures is essential for effective interpretation of sign language. Therefore, while this model excels in recognizing static signs, its utility in real-world scenarios involving the translation of sign language words and sentences, which inherently involve motions and gestures, may be limited.

5.2. Dynamic Signs Recognition Results

For the dynamic sign recognition, we employ a Recurrent Neural Network, as described in Section 4.3.2. The RNN models effectively process sequential data, allowing it to capture the temporal relationships inherent in dynamic gestures, thereby enhancing the accuracy of our recognition system. This capability not only aligns with the linguistic characteristics of sign language but also underscores the importance of utilizing appropriate model architectures to address the unique challenges presented by gesture recognition, ultimately leading to more effective communication solutions for the deaf and hard of hearing community.

The model was trained on 500 collected video clips, each representing one of five distinct Hebrew words, and achieved an accuracy rate of 91%. Table 2 shows precision, recall, and F1-score for each of the trained words. This high level of accuracy underscores the effectiveness of the RNN in handling complex temporal patterns within the video data, such as variations in hand size. This adaptability is crucial, as every individual may have slightly different signing styles and physical attributes, making the model robust and versatile in real-world applications.

Table 2. Classification report for the dynamic words, using RNN.

Class	English equivalent	P	R	F1-score
שלום	hello	0.98	0.97	0.97
תודה	thanks	0.99	1	0.99
להתראות	bye	0.89	0.96	0.92
כן	yes	0.92	0.94	0.93
לא	no	0.87	0.91	0.89
Average		0.95	0.96	0.94

The RNN's ability to maintain high accuracy despite these variations illustrates its strength in capturing the nuances of dynamic gestures. This performance highlights the potential of RNNs in creating more inclusive and accessible communication solutions for the deaf and hard of hearing community.

Overall, our approach demonstrates the importance of selecting the appropriate neural network architecture to address the unique challenges of gesture recognition. By leveraging the strengths of RNNs, we have developed a system that not only aligns with the linguistic characteristics of sign language but also advances the field of accessible technology. This work exemplifies how the thoughtful application of machine learning techniques can significantly impact social inclusivity, providing more effective communication tools and fostering better understanding in diverse communities.

Furthermore, our success with the RNN model opens avenues for future research and development. For instance, expanding the dataset to include a broader range of words and phrases could further enhance the model's performance and applicability. This ongoing research not only contributes

 doi:10.20944/preprints202408.0125.v1

to the technical field but also to the broader goal of making technology more accessible and beneficial to everyone.

5.3. Website Application

The trained model was deployed into a web-based platform-independent application.[5] In this manner, users are able to access the application from any device with an internet connection, regardless of the operating system they use. Figure 5 illustrates three snapshots from the site.

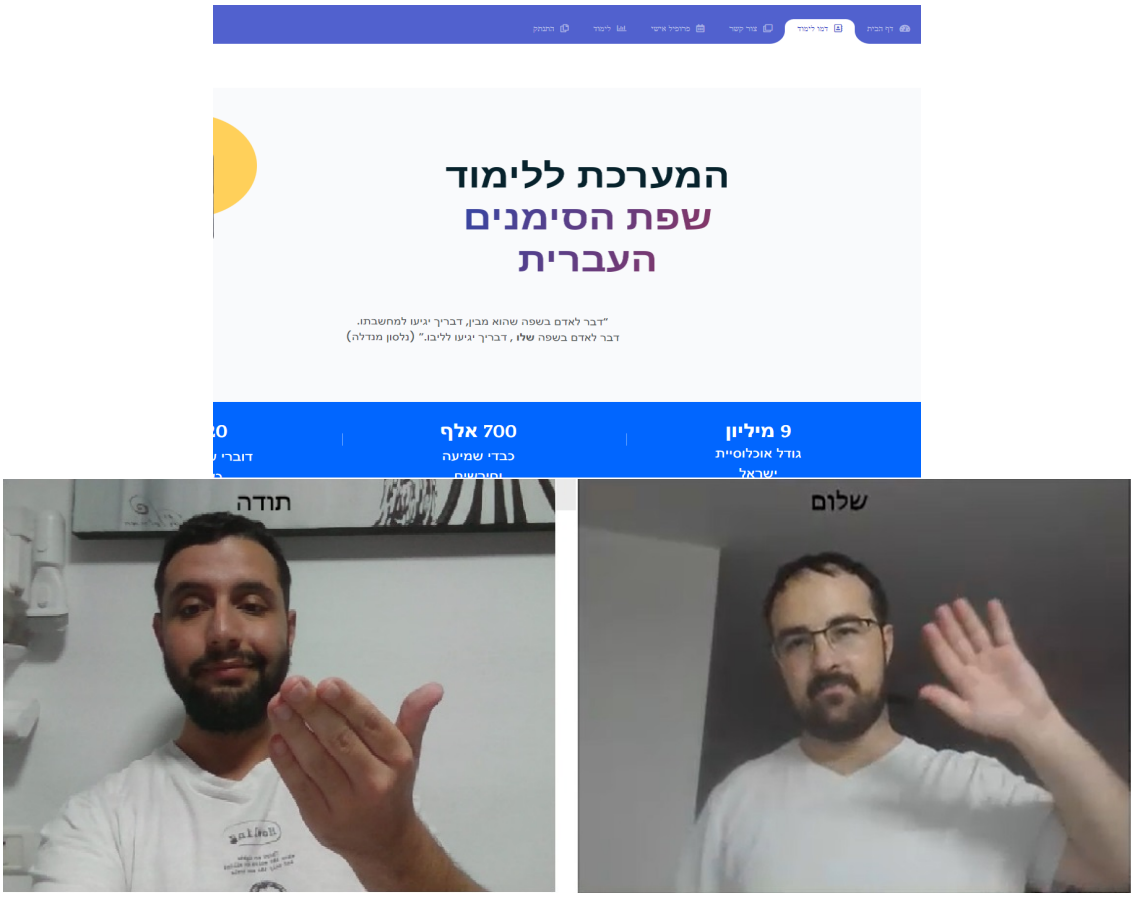


Figure 5. Screenshots from the Signsability website. Top: home page; bottom: examples from the real-recognition model in action. The recognized words are written in the central top of the images.

The developed web-based application puts a strong emphasis on delivering an exceptional user experience, focusing on both intuitive navigation and the effective use of its features. The core function of the website is to provide an intelligent platform for learning sign language. The integration of our RNN model into the website ensures that this learning process is both rapid and precise, seamlessly incorporated into the user interface.

The website is designed with several key elements to enhance usability and engagement:

- **User-Friendly Navigation:** The site features a clear and intuitive layout, allowing users to easily access different sections and functionalities. Menu options and interactive elements are strategically placed to facilitate smooth exploration and learning.

---

[5] URL: http://129.159.153.217.
Development of the site is ongoing; additional features and improvements are constantly added.

 doi:10.20944/preprints202408.0125.v1

- **Responsive Design:** The website is optimized for various devices, including desktops, tablets, and smartphones. This ensures that users can engage with the learning materials from any device, providing flexibility and convenience.
- **Multimedia Integration:** To enhance the learning experience, the site integrates multimedia resources such as instructional videos, animated demonstrations, and example gestures. These resources help users understand and practice sign language more effectively.
- **Personalized Learning Experience:** The website offers personalized features such as progress tracking dashboards. These features are designed to meet individual learning needs and preferences, making the experience more engaging and effective.

By focusing on these aspects, the website aims to provide a comprehensive and user-centric platform for learning sign language, leveraging advanced technology to enhance educational outcomes and accessibility.

6. Conclusions and Future Work

This study presents a comprehensive approach to Israeli Sign Language recognition, leveraging state-of-the-art techniques in video processing and deep learning. The classification pipeline, incorporating MediaPipe for landmark extraction and an RNN model with LSTM layers for sequence classification, demonstrates promising results in translating dynamic gestures into Hebrew words. This project not only underscores the potential of advanced machine learning algorithms in addressing the complex challenge of sign language recognition but also highlights the crucial role of technology in bridging communication gaps for the deaf and hard of hearing community.

The significance of this work lies in its ability to enhance accessibility, making it an important contribution to the field of assistive technologies. The insights gained from this research pave the way for future advancements and applications in sign language recognition, ultimately contributing to a more inclusive society. By continuing to develop the system and extending our dataset, we anticipate it will significantly facilitate the learning of Israeli Sign Language and enhance the accessibility of conversations between hearing and hard of hearing individuals.

We see the promotion of individuals with disabilities as a paramount value, guiding us throughout the project. Encouragement of additional projects related to people with disabilities is crucial for fostering an inclusive and supportive environment.

By prioritizing these values, we aim to not only enhance accessibility and communication for the deaf and hard of hearing community but also inspire further advancements and innovations in assistive technologies. Our commitment to inclusivity drives us to continually seek new ways to improve the quality of life for individuals with disabilities, ensuring they have equal opportunities to participate fully in society.

Author Contributions: Conceptualization, D.E., S.M., and I.R.; methodology, D.E. and S.M.; software, D.E. and S.M.; validation, D.E. and S.M.; formal analysis, D.E., S.M., and I.R.; investigation, D.E., S.M., and I.R.; resources, D.E. and S.M.; writing—original draft preparation, D.E. and S.M.; writing—review and editing, D.E., S.M., and I.R.; supervision, I.R.; project administration, D.E., S.M., and I.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset and the replayed application can be accessed at https://kaggle.com/datasets/fe1ca51ef2c13347756617aedfc611d5699035cdc74ca72e89a357e466f8ebb5 and http://129.159.153.217, respectively.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chengk, K. American Sign Language vs Israeli Sign Language. *StartASL* **2022**, pp. 2–5.

 doi:10.20944/preprints202408.0125.v1

2. Murali, L.; Ramayya, L.; Santosh, V.A. Sign Language Recognition System Using Convolutional Neural Network and ComputerVision. *International Journal of Engineering Innovations in Advanced Technology* **2022**, *4*, 138–141.
3. Obi, Y.; Claudio, K.S.; Budiman, V.M.; Achmad, S.; Kurniawan, A. Sign language recognition system for communicating to people with disabilities. *Procedia Computer Science* **2023**, *216*, 13–20.
4. Balaha, M.M.; El-Kady, S.; Balaha, H.M.; Salama, M.; Emad, E.; Hassan, M.; Saafan, M.M. A vision-based deep learning approach for independent-users Arabic sign language interpretation. *Multimedia Tools and Applications* **2023**, *82*, 6807–6826.
5. Gogoi, J.B.S.D.A.B.A.C.D. Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning. *Elsevier* **2023**, pp. 1386–1393.
6. Shamitha, S.H.; Badarinath, K. Sign Language Recognition Utilizing LSTM And Mediapipe For Dynamic Gestures Of Indian Sign Language. *International Journal for Multidisciplinary Research* **2023**, *5*, 138–152.