Quanta Research
https://journals.tultech.eu/index.php/qr
ISSN: 2806-3279
Volume 2, Issue 1
DOI: https://doi.org/10.15157/qr.2024.2.1.7-13
Received: 15.05.2024; Revised: 18.07.2024; Accepted: 05.08.2024

Technical Paper

A Real-Time Sign Language to Text Conversion System for Enhanced Communication Accessibility

Mansi.N Munde[1]*, Ganesh.G Jadhav[1], Sushma Gunjal[1], Kamlesh.M Mahale[1], Aditya.A Kale[1]

[1]Department of Artificial Intelligence and Data Science, Ajeenkya D.Y Patil School of Engineering, Pune, India

***mansi.munde@dypic.in**

Abstract

The research addresses the problem of converting American Sign Language (ASL) finger spelling into text in real-time, enhancing communication for the deaf and hard of hearing. A convolutional neural network (CNN) is utilized to recognize hand gestures from camera images, focusing on the position and orientation of the hand to create accurate training and testing data. The methodology involves filtering hand images, followed by classification to predict the corresponding sign language characters. The calibrated images are then used to train the CNN model. Key findings demonstrate that the proposed system effectively recognizes ASL finger spelling with high accuracy, offering a valuable tool for improving accessibility in communication. These findings suggest significant potential for further applications in real-time sign language interpretation.

Keywords: Sign Language Recognition, Convolutional Neural Network (CNN), Neural Networks, American Sign Language (ASL).

INTRODUCTION

American Sign Language (ASL) is a vital mode of communication for deaf and hard-of-hearing individuals, relying on gestures, hand movements, facial expressions, and body language to convey meaning. Unlike spoken languages, ASL does not use auditory cues but is a fully developed language system with its own grammar and syntax [1] and [2]. Sign languages are not universal; they vary regionally, adding complexity to communication across different sign language communities [3].

Research in sign language interpretation has become increasingly significant as it aims to bridge the communication gap between deaf or hard-of-hearing individuals and those who do not use sign language. Gesture recognition systems are pivotal in this research area, facilitating natural communication without the need for interpreters [4]. These systems aim to convert ASL gestures into text and speech, enabling smoother interactions between hearing and non-hearing individuals.

Recent advancements in machine learning and computer vision have significantly contributed to this field. For instance, convolutional neural networks (CNNs) have been effectively utilized for sign language recognition, demonstrating their capability in classifying and interpreting complex gestures [5] and [6]. Moreover, the integration of novel vision-based features and machine learning techniques has further enhanced the accuracy and efficiency of sign language recognition systems [7, 8]. Pigou et al. (2015) demonstrated the effectiveness of convolutional neural networks (CNNs) for interpreting sign language gestures from visual data [1]. Building on this, Zaki and Shaheen (2011) introduced a vision-based feature approach that improved recognition accuracy [2]. Mukai et al. (2017) utilized machine learning and classification trees to recognize Japanese fingerspelling, offering insights into advanced recognition methods [3].

Further developments like Bhat's (2022) [4] study on sign language to text conversion and Gupta's project have provided practical applications for interpreting sign language [5] (Figure 1).

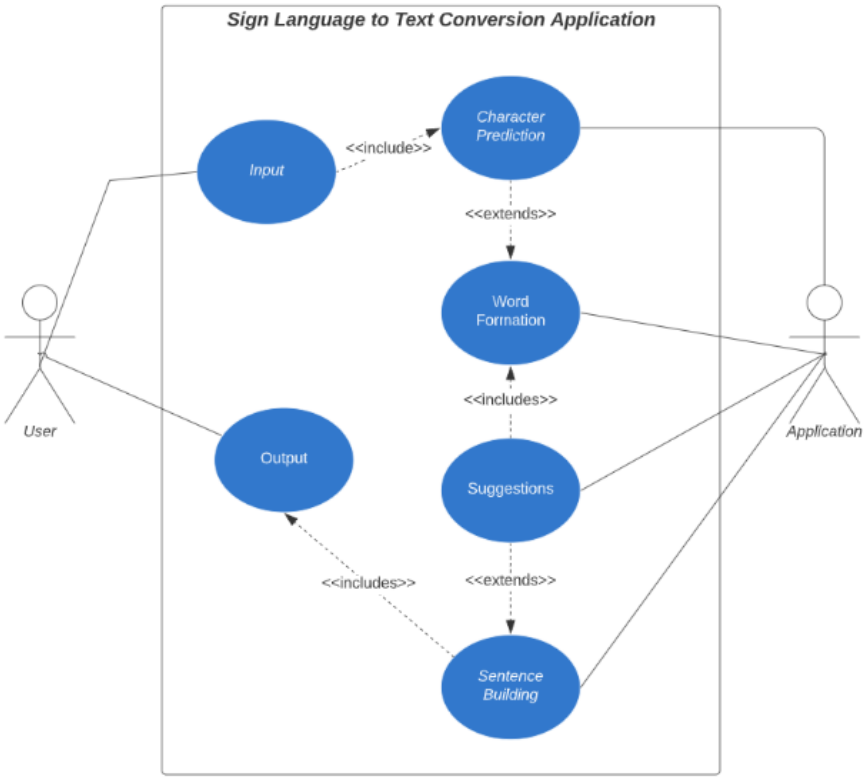

Figure 1. The application working diagram of Gupta's project [5].

Recent studies have expanded these methods: Huang et al. (2015) applied 3D CNNs for enhanced visual recognition [6], Liang et al. (2018) explored real-time recognition with 3D convolutional networks [7], Kanavos et al. (2023) reviewed deep learning advancements in sign language recognition [8], and Zhang et al. (2024) investigated multi-modal recognition using a fusion of deep learning techniques [9]. These studies highlight how technological advancements are improving communication accessibility and bridging gaps between deaf and hearing individuals. Our study focuses on developing a model that recognizes ASL

finger spelling and converts it into text. By leveraging CNNs and combining gesture recognition techniques, we aim to create a robust system that interprets and translates ASL into text and speech, thereby improving accessibility and communication for the hearing impaired.

MATERIALS AND METHODS

The materials used in this research are as follow

1. Extract and display functions: The image is represented as a 3D matrix with height, width, and depth measured by each pixel value (1 for grayscale, 3 for RGB). With CNN, these pixel values are used to extract valuable information.

2. Artificial Neural Network (ANN): ANN can be used to model sequence patterns, extract features from sign language movies, and recognize sign language gestures to text converters. By learning a large set of ANN character movements and accompanying text translations, ANNs improve accuracy and provide real-time translations.

3. Convolutional Neural Network (CNN): Because CNNs can extract spatial features from images or videos, they can be used in linguistic text converters. CNNs help recognize and interpret gestures by accurately identifying hand shapes, gestures, and facial expressions in hand gestures. It can improve the speed and accuracy of text translation, especially for real-time applications.

4. Tensor Flow and Keras: The machine learning framework is available as open source. provide access to tools for model builders and neural trainers. With their help, it is possible to implement CNN and ANN, which translate sign language into text. TensorFlow includes this high-level neural network API. Provides an easy interface to build learning and deep learning models. simplify the process of converting sign language to text for ANN and CNN implementation.

5. OpenCV: A library of freely available computer vision and machine learning software. offers video and image processing tools. Speech to text applications can be used to process and analyse speech video data.

In terms of methods, the system employs a process based on vision. Since all signs are used with the hands alone, the issue of utilizing artificial devices for interaction is resolved and the steps are as follow

1. **Data creation:** Using OpenCV, we were able to obtain about 800 photos for each ASL character for training purposes and about 200 images for testing (as a sample shown in Figure 2), as we were unable to locate an appropriate pre-existing dataset in raw image format. To emphasize important features, we created an area of interest (ROI) for every webcam frame and used a Gaussian blur filter.

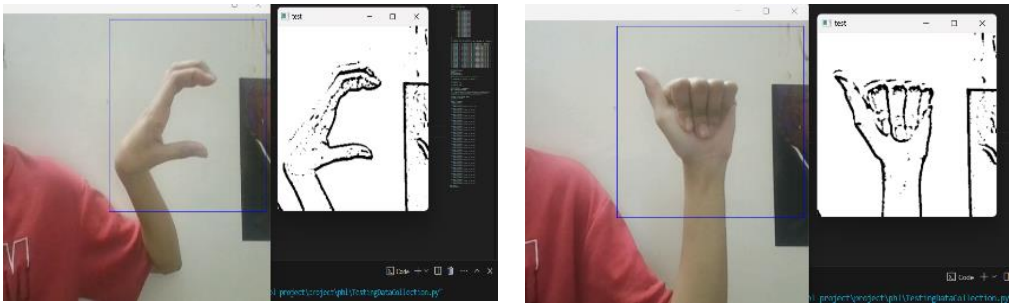

Figure 2. Sample of data creation and gesture classification.

2. **Gesture classification:** We use two tiers of final symbol prediction algorithms in our methodology.

Algorithm level 1: The frame is thresholder and a Gaussian blur filter is applied to extract features. Analyze the picture with a CNN model. Letters that are discovered in more than fifty frames will be examined and printed. The blank character can be used to denote a word break.

Algorithm level 2: Character sets with similarity are identified by detection results. We employ a particular classifier to categorize these collections.

3. **The CNN model:** CNN model is consisting of two layers.

Layer 1: To down sample the image, we use the first layer of convolution and pooling. An additional layer of convolution and pooling processes the image. Densely connected layers are what we utilize for classification. Neurons in the last layer are numbered according to the number of classes.

Layer 2: Apply classifiers to character sets {D, R, U}, {T, K, D, I}, and {S, M, N} that have comparable detection outcomes. This method increases the accuracy of gesture classification, particularly for characters that are visually identical and may cause misunderstanding.

4. **Finger-written sentences avatar:** Print that character and add it to the current line if the number of characters detected beyond a predetermined threshold and there are no nearby characters (e.g. count > 50, threshold = 20). To prevent forecasting the wrong characters, the current character count is cleared if there are nearby characters or if there are enough of them. No spaces are identified if the amount of spaces (normal background) surpasses a particular threshold and the buffer is empty at that moment. If not, the current line is appended to the text below and a space is printed to indicate that the term is about to end.
5. **AutoCorrect features:** we used the Hunspell_suggest Python library to develop the autocorrect feature. The library offers the proper word to input in place of each misspelled one. shows a list of terms that are similar to the word being used, giving the user the option to choose the word and add it to the text. This function lessens spelling mistakes and anticipates challenging words.

6. **Training and Testing:** Convert the input image (RGB) to grayscale, apply Gaussian blur to remove noise, and use adaptive thresholding to extract hands from the background. The image is resized to 128x128. Preprocessed images are input to the model for training and testing. The prediction layer uses the SoftMax function to normalize the output from 0 to 1, making the values of each class sum to 1. Cross-entropy is used as a measure of classification performance and is optimized by adjusting neural weights. A network that uses gradient descent, specifically the Adam algorithm. TensorFlow's optimization tool. The model is trained using labeled data to minimize cross entropy, ensuring that the predicted values are as close as possible to the actual values.

RESULTS AND DISCUSSION

In our study, we evaluated the performance of our sign language recognition system by testing its accuracy across different levels of complexity. When employing only the first level of our system, we achieved a notable accuracy of 95.8%. This baseline performance is quite strong, but we observed a significant improvement when we integrated both levels 1 and 2 of the system, resulting in an impressive accuracy of 98.0%. This improvement demonstrates the robustness of our approach and its potential to outperform several contemporary American Sign Language (ASL) recognition systems reported in the literature. As a discussion, a review of previous research in the field shows a variety of approaches to sign language recognition, often utilizing advanced technologies such as Kinect for hand detection. For example, Pigou et al. [1] developed a sign language recognition system for Flemish sign language that leveraged Convolutional Neural Networks (CNN) in combination with Kinect technology, achieving an error rate of 2.5%. Although this approach is effective, it relies heavily on specific hardware, which may limit its applicability in different settings. Zaki and Shaheen [2] introduced a model focused on a lexicon of 30 words, employing a hidden Markov model to process the signs, which resulted in a higher error rate of 10.90%. This model, while innovative, indicates the challenges faced when expanding the lexicon or transitioning to different signers. Mukai et al. [3] reported their findings on a Japanese Sign Language recognition system, where they achieved an average accuracy of 86% for 41 static gestures. Their system used depth sensors, which provided an accuracy range of approximately 83-85% for new signers and an exceptionally high accuracy of 99.99% for signers who were observed during training. This result underscores the system's dependency on the familiarity of signers to achieve optimal performance. In a different study, Bhat [4] presented an alternative methodology to sign language translation, which further highlights the diversity of approaches within this field. Meanwhile, Huang et al. [6] achieved significant advancements in visual recognition of sign language using CNNs, a popular and powerful tool in image processing and classification tasks. Liang et al. [7] demonstrated the application of 3D convolutional networks for real-time sign language recognition, showcasing the potential of leveraging temporal information to enhance system performance. Their work illustrates the ongoing trend toward more sophisticated and accurate models that can operate in real time. In a

comprehensive review, Kanavos et al. [8] discussed the progress in enhancing sign language recognition using deep learning techniques. Their analysis highlighted the rapid advancements and the increasing capabilities of deep learning models in handling complex recognition tasks. Wang and Zhang [9] explored a multi-modal approach, which combined various deep learning techniques to improve sign language recognition accuracy. Their findings suggest that integrating multiple data modalities can lead to significant improvements in recognition performance, particularly in challenging scenarios where single-modal approaches may fall short.

Overall, our study's results, achieving 98.0% accuracy with a multi-level approach, represent a significant step forward in the field of sign language recognition, particularly when compared to previous studies that have employed a range of methodologies and technologies. Unlike many existing models that incorporate background subtraction to improve accuracy, our model omits this step, focusing instead on affordability and accessibility. By utilizing standard webcams rather than specialized devices like Kinect, our project offers a more cost-effective and widely available solution, with potential accuracy trade-offs in certain conditions.

CONCLUSION AND FUTURE RESEARCH

Real-time feature recognition for D&M (deaf and mute) individuals in American Sign Language is created in this study for the ASL alphabet. With our dataset, we ultimately obtained a 98.0% accuracy rate. After adding two layers to the algorithm that determined which characters were more similar to one another, we were able to make more accurate predictions. If there is enough brightness, no background noise, and nearly all characters are displayed correctly, this makes it possible to detect virtually all of the characters. With CNN and ANN, the Sign Language Converter has a solid base upon which to grow in the future. Enhancing gesture recognition with larger datasets and better architectures, as well as optimizing real-time translation for smooth communication, are important areas for progress. Expanding one's vocabulary can be accomplished by either improving models or growing the dataset. Accuracy can be improved through multimodal integration by using body language and facial emotions. Creating wearable and mobile applications can increase accessibility. Enhancements to the user interface that are beneficial to both signers and non-signers include intuitive feedback. The system may be adjusted for various sign languages and geographical areas through localization and customization. Assistive technology integration, such as text-to-speech, can result in all-encompassing communication solutions. These improvements have the potential to greatly increase the deaf and hard-of-hearing community's accessibility to communication.

CONFLICT OF INTERESTS

The authors confirms that they have no conflicts of interest related to the publication of this research.

REFERENCES

1. Pigou, L.; Dieleman, S.; Kindermans, P.-J.; Schrauwen, B. Sign Language Recognition Using Convolutional Neural Networks. In Computer Vision - ECCV 2014 Workshops; Agapito, L., Bronstein, M.M., Rother, C., Eds.; Springer International Publishing: Cham, 2015; Vol. 8925, pp. 572–578 ISBN 9783319161778.
2. Zaki, M.M.; Shaheen, S.I. Sign Language Recognition Using a Combination of New Vision Based Features. Pattern Recognition Letters 2011, 32, 572–577, doi:10.1016/j.patrec.2010.11.013.
3. Mukai, N.; Harada, N.; Chang, Y. Japanese Fingerspelling Recognition Based on Classification Tree and Machine Learning. In Proceedings of the 2017 Nicograph International (NicoInt); IEEE: Kyoto, Japan, June 2017; pp. 19–24.
4. Bhat, A.; Yadav, V.; Dargan, V.; Yash Sign Language to Text Conversion Using Deep Learning. In Proceedings of the 2022 3rd International Conference for Emerging Technology (INCET); IEEE: Belgaum, India, May 27 2022; pp. 1–7.
5. Gupta, Nikhil. "Sign Language to Text Conversion." GitHub, 29 Oct. 2023, github.com/emnikhil/Sign-Language-To-Text-Conversion.
6. Jie Huang; Wengang Zhou; Houqiang Li; Weiping Li Sign Language Recognition Using 3D Convolutional Neural Networks. In Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME); IEEE: Turin, Italy, June 2015; pp. 1–6.
7. Liang, Z.; Liao, S.; Hu, B. 3D Convolutional Neural Networks for Dynamic Sign Language Recognition. The Computer Journal 2018, 61, 1724–1736, doi:10.1093/comjnl/bxy049.
8. Kanavos, A.; Papadimitriou, O.; Mylonas, P.; Maragoudakis, M. Enhancing Sign Language Recognition Using Deep Convolutional Neural Networks. In Proceedings of the 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA); IEEE: Volos, Greece, July 10 2023; pp. 1–4.
9. Zhang, P.; Wang, D.; Lu, H. Multi-Modal Visual Tracking: Review and Experimental Comparison. Comp. Visual Media 2024, 10, 193–214, doi:10.1007/s41095-023-0345-5.