Sign Language to Text Conversion using CNN

Alan Wilson, Lenet Steephen

***Abstract:** Sign language is a communication strategy used by those who are unable to hear. So those people who know sign language can communicate with people who are deaf. But a majority of our people don't know sign language therefore there comes a communication gap between the ones who know sign language and others who don't know. This project's major purpose is to bridge this gap by developing a system that recognizes multiple sign languages and translates them into text in real-time. We use machine learning technologies to construct this system especially, convolutional neural networks (cnns), which are used to recognize and translate American Sign Language (ASL) into text by capturing it using a webcam. The transformed text is then presented on the screen by which individuals can comprehend and communicate with those who use sign language. The system's performance is evaluated on a dataset of ASL gestures, attaining excellent accuracy and indicating its potential for practical usage in enhancing communication accessibility for the deaf and hard-of-hearing community.*

***Keywords:** Sign Language, Convolutional Neural Network (CNN), Real-time, American Sign Language (ASL)*

I. INTRODUCTION

Sign language is a communication strategy in which we use our hands by showing various gestures that are utilized by persons who are deaf or hard of hearing to engage with each other. Sign language is a demonstrative way of communication. But there is a difficulty when interacting with those who don't know sign language, therefore it is difficult to grasp, especially when there is no sign language interpreter available. This causes people with hearing challenges to completely engage in society, including educational, professional, and social situations.

To solve this situation I have developed a system that can convert hand gestures into written words. Gestures are body movements that express an idea or something into a meaning. This is usually known as sign language recognition systems. To evaluate and understand the motions performed by a person, computer vision and machine learning techniques are commonly used.The system can produce written words by detecting the movements captured using a webcam. Thus it clears the way for the interaction between people who use sign language and those who don't. The main objective of this project is to develop a system that accurately translates American Sign Language (ASL) gestures into English written text.

Manuscript received on 16 April 2024 | Revised Manuscript received on 04 May 2024 | Manuscript Accepted on 15 May 2024 | Manuscript published on 30 May 2024.
**Correspondence Author(s)*

Alan Wilson*, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. E-mail: alanwilson476@gmail.com, ORCID ID: 0009-0006-4347-5868

Lenet Steephen, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. E-mail: lenet@alberts.edu.in.

To identify and understand American Sign Language (ASL) in real-time camera recording, I have used convolutional neural networks (CNNs) which is a deep learning methodology. Thus a user interface is developed and the translated text is shown on that screen instantly and here which smoothens the communication

We have created this system to assist people who are deaf or hard of hearing. By this, we have made a valuable endowment to the progress of assistive technology. The ultimate aim is to improve their communication skills and help them to engage more actively in society. This project aims to satisfy the requirement for better communication between deaf people and others. It is in line with the ideals of inclusion and technology. It is in keeping with the concepts of inclusivity and technology

Through the creation of this system, our goal is to make a valuable contribution to the progress of assistive technology for individuals who are deaf or hard-of-hearing. Ultimately, we aim to improve their communication skills and enable them to actively engage in a society that primarily relies on hearing. This project seeks to satisfy the urgent requirement for better interaction and usability between people who are deaf or hard of hearing. It is in line with the ideals of inclusion and technology

II. LITERATURE REVIEW

[1] Pujan Ziaie, Thomas M¨uller, Mary Ellen Foster, and Alois Knoll, presented an effective and fast method for static hand gesture recognition. This method is based on classifying the different gestures according to geometric-based invariants which are obtained from image data after segmentation; thus, unlike many other recognition methods, this method is not dependent on skin color. Gestures are extracted from each frame of the video, with a static background. The segmentation is done by dynamic extraction of background pixels according to the histogram of each image. Gestures are classified using a weighted K-Nearest Neighbors Algorithm which is combined with a nave Bayes approach to estimate the probability of each gesture type.

[2] Mohammed Waleed Kalous et al. In this project, Instrumented gloves use a variety of sensors to provide information about the user's hand. They can be used for recognition of gestures; especially well-defined gesture sets such as sign languages. However, recognizing gestures is a difficult task, due to intrapersonal and interpersonal variations in performing them. One approach to solving this problem is to use machine learning. In this case, samples of 95 discrete Australian Sign Language (Auslan) signs were collected using Power-Glove. Two machine learning techniques were applied {instance-based learning (IBL) and decision-tree learning {to the data after some simple features were extracted.

Accuracy of approximately 80 percent was achieved using IBL, despite the severe limitations of the glove

[3][6][7][8][9][10] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans & Benjamin Schrauwen et al. There is an undeniable communication problem between the Deaf community and the hearing majority. Innovations in automatic sign language recognition try to tear down this communication barrier. Our contribution considers a recognition system using the Microsoft Kinect, convolutional neural networks (CNNs) and GPU acceleration. Instead of constructing complex handcrafted features, CNNs are able to automate the process of feature construction. We are able to recognize 20 Italian gestures with high accuracy. The predictive model is able to generalize on users and surroundings not occurring during training with a cross-validation accuracy of 91.7%. Our model achieves a mean Jaccard Index of 0.789 in the ChaLearn 2014 Looking at People gesture spotting competition.

[4] Zaki, M.M., Shaheen, S.I. et al. Sign languages are based on four components hand shape, place of articulation, hand orientation, and movement. This paper presents a novel combination of vision based features in order to enhance the recognition of underlying signs. Three features are selected to be mapped to these four components. Two of these features are newly introduced for American sign language recognition: kurtosis position and principal component analysis, PCA. Although PCA has been used before in sign a language as a dimensionality reduction technique, it is used here as a descriptor that represents a global image feature to provide a measure for hand configuration and hand orientation. Kurtosis position is used as a local feature for measuring edges and reflecting the place of articulation recognition. The third feature is motion chain code that represents the hand movement. On the basis of these features a prototype is designed, constructed and its performance is evaluated. It consists of skin color detector, connected component locator and dominant hand tracker, feature extractor and a Hidden Markov Model classifier. The input to the system is a sign from RWTH-BOSTON-50 database and the output is the corresponding word with a recognition error rate of 10.90%.

[5] Byeongkeun Kang, Subarna Tripathi, Truong Q. Nguyen et al. Sign language recognition is important for natural and convenient communication between deaf community and hearing majority. We take the highly efficient initial step of automatic fingerspelling recognition system using convolutional neural networks (CNNs) from depth maps. In this work, we consider relatively larger number of classes compared with the previous literature. We train CNNs for the classification of 31 alphabets and numbers using a subset of collected depth data from multiple subjects. While using different learning configurations, such as hyper-parameter selection with and without validation, we achieve 99.99% accuracy for observed signers and 83.58% to 85.49% accuracy for new signers. The result shows that accuracy improves as we include more data from different subjects during training. The processing time is 3 ms for the prediction of a single image. To the best of our knowledge, the system achieves the highest accuracy and speed. The trained model and dataset is available on our repository1.

III. METHODS

Gathering information is an essential component of sign language-to-text conversion. For this work, I attempted to locate pre-existing datasets, however was unable to identify any datasets consisting of unprocessed photos that met the particular needs. The only datasets that were able to be located contained the format of RGB values. So decided to establish my dataset. The process of creating a dataset for sign language-to-text conversion entails recording gestures in sign language together with their accompanying textual identifiers. The technique utilizes a vision-based methodology. Every sign is conveyed using only the hands, therefore obviating the need for any manmade tools in communication. The process I employed to generate my dataset was by using Open Computer Vision (OpenCV) library to generate my dataset. Recorded using a high-resolution webcam or camera, capturing a range of sign language gestures. Ensure adequate illumination and an unobstructed view of the signer's hands. Initially, I obtained approximately 800 photographs of each symbol in ASL (American Sign Language) for the sake of training, and around 200 images per symbol for testing. Initially, I got each frame displayed by my device's webcam. Within each frame, I establish a Region of Interest (ROI) that is visually represented by the shape of a square with a blue boundary. Next, utilize the Gaussian Blur Filter on the picture to extract different characteristics of the image that reduce the noise on the image that had captured through the camera

Utilize the Gaussian Blur filter and thresholding technique on the frame captured with openCV to obtain the resultant picture following feature extraction. The decoded picture is inputted into the Convolutional Neural Network (CNN) algorithm to perform prediction. If a letter is identified in more than 50 consecutive frames, it is printed and considered in the formation of a word. The presence of space among words is denoted by the blank symbol. I identify multiple sets of symbols that exhibit comparable detection outcomes. Subsequently, I employ specialized classifiers to categorize the data into their respective sets.

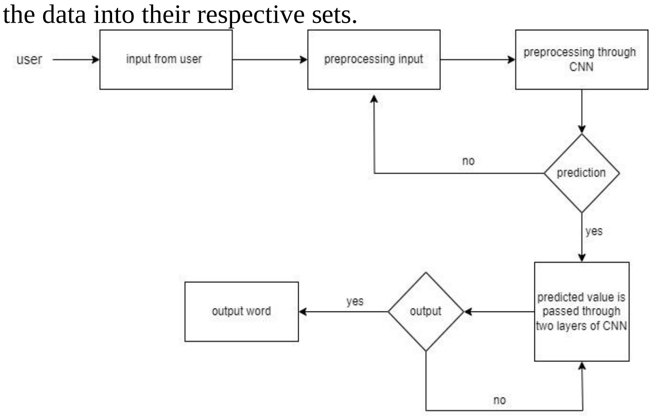
user
input from user
preprocessing input
preprocessing through CNN
prediction
no
yes
predicted value is passed through two layers of CNN
output
yes
output word
no

Fig.1 Gesture Classification

I have employed a dual-tier strategy, comprising a Convolutional Neural Network (CNN) model with two convolution layers as the initial layer.

The input image has a resolution of 128x128 pixels and undergoes processing by applying 32 filter weights, each measuring 3x3 pixels. The photos are reduced in size using a 2x2 max pooling approach. Our picture's resolution has been decreased to 63x63 pixels using down-sampling. The input data for the second convolutional layer is obtained from the output of the first pooling layer, which comprises 63x63 units. The second convolutional layer utilizes 32 filter weights, with each weight measuring 3x3 pixels, to process the input. This will result in a picture with dimensions of 60 pixels by 60 pixels. The resultant photographs are subjected to a further down-sampling process using a 2x2 max pool operation, resulting in images with a resolution of 30x30.

The images are being employed as input for a fully linked layer comprising 128 neurons. The result obtained from the second convolutional layer is subsequently converted into an array consisting of 30x30x32 = 28800 elements. The input to this layer comprises an array containing 28,800 items. The output of these layers is subsequently utilized as the input for the second Densely Connected Layer.

The output obtained from the initial Densely Connected Layer is used as the input for a fully connected layer of 96 neurons. The output of the second Densely Connected Layer is used as the input for the final layer, which will have the same number of neurons as the number of classes being used. In the following algorithm, I utilize algorithms to validate and forecast symbols that exhibit a higher degree of similarity to one another, enabling me to approach the detection of the displayed sign as closely as possible.

I am using two layers of the algorithm to verify and predict similar symbols so that can detect the symbols shown. During testing, I discovered that certain symbols were not displaying correctly and producing unintended symbols. D represents the relationship between R and U, meaning that R is dependent on U. Similarly, U represents the relationship between D and R, indicating that D is dependent on R. My initials are T, D, K, and I. Regarding S, the relationship between M and N is being discussed. To address the aforementioned scenarios, we have developed three distinct classifiers to categorize the following sets: {D, R, U}, {T, K, D, I}, and {S, M, N}. Thus the symbol is detected by the system using CNN and the predicted word is shown on the Interface. By combining the predicted alphabets the words are formed. I used a python library named Hunspell to suggest the words while predicting the alphabet. So we can auto-correct the word directly.

IV. RESULTS AND DISCUSSION

A. Results

The model that I had developed achieved an accuracy of 95.8% by exclusively utilizing layer 1 of our algorithm. By integrating both layer 1 and layer 2, I have attained a precision of 98.0%, surpassing the precision of the latest scholarly articles on American sign language. Most research papers focus mainly on the utilization of devices like Kinect for the specific purpose of detecting hands. Also, there are different methods of using kinetic devices for detecting motions. By using a hidden Markov model classifier they have achieved an error rate of 10%. Also by using depth sensors, they have achieved an error of 9.1% error rate

B. Discussion

1. Implications: The potential of this project is that it provides the ability to communicate between sign language users and non-users by accurately translating gestures into text. So they can communicate easily and effectively. This allows us to improve the availability of educational resources for learners with hearing impairments in real-time. By using deep learning methods like convolutional neural networks (CNNs), the project shows the practicality and effectiveness of these methods.

2. Challenges and Limitations: The primary challenge in this project was obtaining a dataset, as there were no pre-existing datasets available. So I decided to create my dataset. So I decided to utilize convolutional neural networks (CNN) in Keras to analyze raw images that are square-shaped. This choice was motivated by the enhanced ease and efficiency of working exclusively with square images. The next challenge was to find an appropriate filter to extract the relevant features from the captured images. After trying various filters like binary threshold, canny edge detection, and the Gaussian blur filter, I finally decided to make use of the Gaussian blur filter. Then those images that are filtered are passed to the CNN model for prediction. I also faced issues related to the accuracy of the model that had already been trained, and to solve this situation, I increased the dimensions of the input image, therefore improving the dataset.

3. Future Directions: The current system could predict words on a white background, so I aim to improve the precision while working with various complex backgrounds. This can be done by testing various background subtraction algorithms. Also to improve the accuracy in predicting the gestures in low light. They make this system more useful by considering developing a mobile application so that people can use it more comfortably. The current project mainly focuses on American sign language prediction, but in the future, this system can be extended to additional sign languages through the collection of sufficient data and the implementation of suitable training.

V. CONCLUSION

In conclusion, the development of a Convolutional Neural Network (CNN) model for translating sign language into written text represents a significant advancement in the realm of assistive technology and enhancing accessibility. This project has effectively employed deep learning techniques to showcase the feasibility of automatically translating sign language gestures into text, thereby improving communication for individuals with hearing impairments. This report details the creation of a practical, real-time vision-based system that can recognize signs in American Sign Language (ASL). The system is designed to assist individuals who have impairments in hearing and speech. The system is designed to specifically identify and distinguish American Sign Language (ASL) alphabets.

I have achieved a definitive accuracy of 98.0% on our dataset. I have improved our prediction by incorporating two levels of algorithms that validate and forecast symbols that demonstrate a higher degree of resemblance to each other.

This enables us to accurately recognize almost all symbols, provided that they are presented correctly, there is no disruption from background noise, and the lighting is adequate.

DECLARATION STATEMENT

Funding	No, I did not receive
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Yes, It is relavant. The dataset used for this project was Grid Corpus Dataset GRID is an openly available corpus containing an audio-visual database from 34 speakers with 1000 utterances per speaker.
Authors Contributions	Each author has made an independent contribution to the article. The individual contributions of each author are presented below for clarity and transparency. Alan Wilson is the main contributor and Ms. Lenet Steephen is the project guide.

REFERENCES

1. Pujan Ziaie, Thomas M uller, Mary Ellen Foster, and Alois Knoll “A Na ̈ive Bayes Munich, Dept. of Informatics VI, Robotics and Embedded Systems, Boltzmannstr. 3, DE-85748 Garching, Germany.
2. Mohammed Waleed Kalous, Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language.
3. Pigou L., Dieleman S., Kindermans PJ., Schrauwen B. (2015) Sign Language Recognition Using Convolutional Neural Networks. In: Agapito L., Bronstein M., Rother C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science, vol 8925. Springer, Cham
4. Zaki, M.M., Shaheen, S.I.: Sign language recognition using a combination of new vision based features. Pattern Recognition Letters 32(4), 572–577 (2011).
5. Byeongkeun Kang, Subarna Tripathi, Truong Q. Nguyen” Real-time sign language fingerspelling recognition using convolutional neural networks from depth map” 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR).
6. Radhamani, V., & Dalin, G. (2019). Significance of Artificial Intelligence and Machine Learning Techniques in Smart Cloud Computing: A Review. In International Journal of Soft Computing and Engineering (Vol. 9, Issue 3, pp. 1–7). https://doi.org/10.35940/ijsce.c3265.099319
7. Netay, I. V. (2022). Influence of Digital Fluctuations on Behavior of Neural Networks. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 3, Issue 1, pp. 1–7). https://doi.org/10.54105/ijainn.a1061.123122
8. A., O., & O, B. (2020). An Iris Recognition and Detection System Implementation. In International Journal of Inventive Engineering and Sciences (Vol. 5, Issue 8, pp. 8–10). https://doi.org/10.35940/ijies.h0958.025820
9. Kaur, J., & Gupta, N. (2019). Constructive Neural Network: A Framework. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 2, pp. 5321–5324). https://doi.org/10.35940/ijeat.b3304.129219
10. Magapu, H., Krishna Sai, M. R., & Goteti, B. (2024). Human Deep Neural Networks with Artificial Intelligence and Mathematical Formulas. In International Journal of Emerging Science and Engineering (Vol. 12, Issue 4, pp. 1–2). https://doi.org/10.35940/ijese.c9803.12040324

AUTHORS PROFILE

Alan Wilson, currently pursuing Master of Science in Computer Science from the prestigious St. Albert’s College (Autonomous), Ernakulam. Prior to this he had completed his Bachelor of Science degree in Computer Science from KMM College, Ernakulam. His area of interests includes prominent fields like Networking, Database Management,IOT. He is given attention to details as well as he is able to think outside the box, he loves to solve problems and has been keenly observing the latest technology. He is an active member of the Computer Science community.

Ms. Lenet Steephen is an Assistant Professor at St. Albert's College (Autonomous) with a strong academic background and industry experience. She completed her undergraduate studies at St. Albert's College (Autonomous), Ernakulam, followed by postgraduate studies at the School of Computer Sciences, Mahatma Gandhi University, Kottayam. Prior to her current role, Ms. Steephen worked as a Programmer at Tata Consultancy Services, Chennai, gaining valuable insights and practical skills for over two years. Her blend of theoretical knowledge and hands-on experience makes her a valuable asset in the field of computer science education and research.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.