Real Time Sign Language Recognition

Pankaj Kumar Varshney (pankaj.surir@gmail.com)
Institute of information Technology and Management GGSIPU New Delhi

Gaurav Kumar
Sharda University

Shrawan Kumar
Shoolini University

Bharti Thakur
Shoolini University

Plakshi Saini
Institute of information Technology and Management GGSIPU New Delhi

Vanshika Mahajan
Institute of information Technology and Management GGSIPU New Delhi

---

Research Article

Keywords: Machine Learning, Convolution Neural Network (CNN), Sign Language Recognition, American Sign Language (ASL)

Posted Date: May 11th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2910431/v1

Abstract

Speaking with someone who have hearing loss may be quite challenging. Systems that can recognize different signs and alert regular people are thus required. Recognition of sign language is a big development in assisting deaf-mute persons. With the exception of J and Z, which require motion detection for recognition, the objective of this study is to create a model based on neural networks for precise and user-friendly sign language identification that can identify finger spelling-based hand gestures representing the ASL alphabets.

1. INTRODUCTİON

Each of us has a certain communication style to interact with other people and comprehend the environment. Although speaking is the most prevalent mode of communication, there are many other means, such as body language, gestures, reading and writing, and using visual aids. Sadly, there is still a communication gap for the small group of people who have speech and hearing difficulties.

These individuals rely only on sign language to communicate with others in their social context. Visual cues include the hands, eyes, face, lips, and body. The actions or symbols are grouped linguistically in sign language. It incorporates a range of nonverbal communication strategies, including body language, facial expressions, timing, touch, and anything else.

As we have discussed before, the domain of sign language based on recognizing hand posture and position for human-computer interaction is a part of image-based research where a machine learning approach can be utilized to identify the intended meaning of sign gestures. This would allow machines to accurately interpret the signs posed by individuals who are deaf and mute. Thereby, the main aim of the research can be to focus on the gestures and position of the hand to deliver the information or metadata that can be interpreted to control a device or reflect the outcome of a device for a specific requirement. The spatial domain can be considered for the modeling of gestures, whereas postures with dynamic movement can be focused for identification under the area of the temporal domain. The approach for identification of gesture recognition can be considered with the help of two approaches: either a vision-based approach or the dynamic movement of the hand can be captured with the help of a data glove.

The research article consider the vision based approach for creating a system that could be able to identify the actual requirement based upon real time processing. The vision approach is comparatively simpler to be materialized in comparison to glove based approach and provide instinctive approach for human computer interaction.

The hand gesture certainly reflect an important part in day-to-day life which is main aim of this article to address but at the same time there are various domains where the application can be extended like touch screen, gaming consoles, medical imaging, augmented reality etc.

The sign language is a general and effective way for people with hearing and voice impairments to communicate, but it's apparent that the commerce between normal mortals and people with the signaled disabilities becomes delicate in general. It generally leads to difficulty in communication. In order to address the issue of interference, it is essential that sign language is built upon a comprehensive vocabulary of signs, much like spoken language is based on a collection of words. The major challenge bone can face in sign language recognition is that all over the world there are various types of languages and different hand movements subscribe languages that aren't formalized and common, and there is further variation in the alphabet at different geographic locations. There are colorful signs and movements associated with Portuguese sign language (1,2). Sign Language Recognition (SLR) systems aim to offer an accurate and effective means of translating sign language into written or spoken language for a range of applications, including facilitating computer-based interactions for young children through sign language comprehension. Because SLR suggests using hand gestures to communicate important ideas (2), conservative characteristics at birth and selection are pivotal factors to take into account. As visual characteristics describe the substance of the image (3), the unborn effectiveness of the recognition system depends on their selection of the applicable image categorization system. Stoner independence and standpoint invariance are two pivotal characteristics for these kinds of systems. This study presents and describes a vision-grounded system that can crack stationary hand movements in the Portuguese alphabet. Rapid Miner was used to assay the chosen hand features. (4). The entire structure of this composition is grounded in the literature review, and for added complexity, a neural network-based approach is used for the identification of sign language.

2. RELATED WORK (LITERATURE REVIEW)

Stationary or dynamic hand gesture discovery for real-time systems involving mortal computer commerce is a delicate task that's presently the subject of active exploration. There are multitudinous well-presented explorations on gesture recognition and disquisition done in the given compositions (5, 6). Hand gesture recognition and segmentation are crucial tasks in computer vision applications that involve real-time human-computer interaction. The outgrowth of this stage will determine whether a particular hand shape matches a certain model in the future or whether the representative class has the most parallels. As Wacs notes (p. 7), the proper selection of points and their integration with advanced recognition and classification algorithms play a crucial role in determining the success or failure of current and future projects in the field of human-computer interaction through hand gestures.

While generalized machine learning techniques are commonly used for various classification tasks, such as spam classification utilizing the TF-IDF approach [8], Trigueiros performed a comparative analysis of seven distinct algorithms for extracting hand features with the specific objective of categorizing static hand gestures [9].Trigueiros' study revealed that, in terms of computational complexity, utilizing the radial signature and centroid distance as independent features produced the best results for static hand gesture classification. Additionally, he has developed a vision-based system [10] that enables a wheelchair to be controlled with fewer finger commands than typically required. The system detects and segments the user's hand, extracting the fingertips as features to generate commands for controlling the wheelchair.

Wang [11] utilized the discrete Adaboost learning algorithm along with SIFT features to attain in-plane rotation invariance, scale invariance, and multi-view hand detection in a vision-based methodology., Conceil [12] compared two distinct shape descriptors, Fourier descriptors and Hu moments, to recognize 11 hand postures. Their findings indicated that Fourier descriptors had better recognition rates compared to Hu moments. Barczak [13] assessed the effectiveness of Fourier descriptors and geometric moment invariants in recognizing American Sign Language gestures from a database. The findings demonstrated that some classes in the database cannot be distinguished using either descriptor. Bourennane [14] conducted a study comparing various shape descriptors for real-time video-based finger gesture recognition, with the aim of achieving a balance between recognition accuracy and computational efficiency. They evaluated two sets of region-based moments and two categories of contour-based Fourier descriptors, which are insensitive to hand translation, rotation, or scale changes. The authors performed extensive tests on their own dataset under realistic conditions, as well as on the Triesch benchmark database [15].

Since it is seen a lot of potential in the application of such a recognition system to handheld devices, a requirement for a solution with a low computational cost was there. The estimation of the body's pose in the form of two-dimensional landmark locations serves as our foundation for recognition. MatyášBoháček, Marek Hrúz present a robust pose normalization method that processes hand poses independently of body poses in a separate local coordinate system while taking into account the signing space.[16]

A public significant task known as nonstop sign language recognition(cSLR) converts a sign language videotape into an ordered buff sequence. Since there's no unequivocal alignment between sign videotape frames and corresponding facades, it's essential to capture the fine- granulated buff- position details.(17).

Hand spelling is an essential component of sign language, as it allows for the communication of names, addresses, and other words that do not have a specific sign associated with them. Despite this, finger spelling is not widely used because it can be challenging to learn and cumbersome to use. Additionally, there is no universal sign language, and it is known by a limited number of individuals, which makes it a less practical option for communication.

A method for categorizing finger spelling in sign language can overcome this problem. The accuracy of multiple machine learning methods used in this work is tracked and compared.

3. FUNCTİONALMODULES

The proposed model consists of several functional modules, beginning with mortal disguise discovery. The first step in this module is to identify the hand disguise and recognize the gesture being made. Next, point birth is crucial, sign languages communicate information through diverse means, including shape of hand, upper body motion, facial expressions, and hand motions. Language recognition techniques are used to extract elements that portray each of these separate aspects of the signs. Finally, the recognition

of mortal disguise is considered, where the mortal disguise is identified, and the intended meaning is understood and interpreted to produce the final output of the functional module.

4. ALGORITHM AND DATASET PROCESSING

The "Sign Language Gesture Images Dataset" by Ahmed Khan on Kaggle has been used for training our model. The original dataset contained 37 different orders, each with 500 images. It consists of Figs. 0 through 9 and all 26 rudiments in addition to the space sign.

However, in the exploration paper training model, only the characters are considered, with the exception of figures such as J and Z, which require additional discovery.

Also, we added a' nothing order to the dataset. Eventually, the outgrowth is considered with 37500 images belonging to 25 different classes. OpenCV and Python are used for data processing.

Convolution Neural Networks (CNN) are a special type of deep neural networks commonly used for image recognition and computer vision tasks. They are designed to effectively identify patterns in image data with a grid-like topology by performing convolutions, which involves applying a filter or kernel to the input image and sliding it over the entire image to extract features. The convolution operation generates a feature map, which is a summary of the image content that highlights the most relevant and distinctive features.

A CNN model consists of four main missions.

- Convolution: CNNs extract features from images by performing convolutions, which involve sliding a filter over the image to extract patterns. After convolution, an activation function such as ReLU is used to introduce non-linearity into the output, enabling the network to learn more complex patterns in the data.
- Relu: The ReLU activation function in a convolution neural network replaces negative pixel values with zero, introducing non-linearity to the network and enhancing its capability to learn and identify patterns in image data.
- Pooling: Pooling is a technique used in CNNs for down-sampling feature maps, which reduces the dimensionality of the data while preserving important features. It involves partitioning the input image into small sections and then taking the maximum, minimum, or average value of each section to obtain a lower-resolution version of the feature map.
- Fully-connected layer: Fully connected layer of CNN is a neural network layer in which each neuron in the layer is connected to every neuron in the previous layer, allowing for complex feature representations and classification.

5. DEPENDENCIES USED

- Tensorflow

Tensorflow, a free and open-source software archive, is used to test differentiable computing and dataflow across a range of conditions. It is a symbol of calculation archives that are also utilised by neural network processes in engine literacy. In both research and production, Google employs it.

- Keras

Keras is an open-source neural network library written in Python. It is designed to simplify the creation of deep learning models and enables fast experimentation by providing a user-friendly interface and pre-built models. Keras is built on top of other machine learning frameworks like TensorFlow and Theano, and it allows users to easily switch between these frameworks as needed.

- Opencv

OpenCV is an open-source computer vision library that provides various tools and functions for real-time computer vision and image processing tasks, including object detection, tracking, segmentation, and more. It is widely used in the field of computer vision and machine learning for developing real-world applications.

- Numpy:

NumPy is a Python library used for scientific computing, particularly for numerical computations with large, multi-dimensional arrays and matrices. It provides various mathematical functions for linear algebra, Fourier transforms, and more, making it a valuable tool for machine learning and data analysis.

6. METHODOLOGY

- Pre-Processing

The pre-processing contains two steps:

Segmentation: Frame-by-frame segmentation of the video. The division of a digital picture into many pieces is known as segmentation. The goal of segmentation is to make an image representation more understandable and straightforward to analyze.

Binarization: Algorithms are used to binarize each and every grayscale image, and they should perform well for photos with intricate backgrounds.

- Feature Extraction

Extraction of features to represent each distinct characteristic of the signs is a very important stage in sign language identification. It might be about the dynamic physical movement of the hands, face, fingers, or entire body. The forms, movements, and textures of the actual hand gestures are very diverse. The feature must be effective and dependable to handle the diversity of these variances.

- Pattern Matching/Recognition

Using the database to analyze successive movements and recognize the commands or behaviours of users, pattern matching and recognition take place at this step.

7. CONCLUSION & RESULT

Using the Python and machine learning technologies we learned, we successfully achieved the goal of our paper, which was to recognize sign language motions in real time. Our accuracy rate came out at 98.83%.

Classification accuracy and loss serve as our main performance measurement indicator for our model. As previously stated, our model achieved an accuracy of 98.83%, and we plotted two learning curves to illustrate the evolution of both accuracy and loss by contrasting both accuracy and loss with the number of epochs. The learning curves we acquired are displayed in the following charts.

The following pictures show the outputs were corded while testing there all-time sign language recognition tool:

8. FUTURE SCOPE

The state of technology changes with time. The ability of this system to adapt to technological advancements is crucial.

In future we can include the following enhancements:

Camera devices may effectively boost hand item identification by using a better system and surroundings, as well as give an accurate and efficient means to translate sign language into text or voice format.

To recognize ASL words and phrases, a model can be developed. For this, a system that can detect changes in temporal space will be required.

To help individuals who struggle with speech and hearing problems and close the communication gap, we may provide a comprehensive offering.

Declarations

No conflict exists:

1. Author Pankaj Kumar Varshney declares that he has no conflict of interest.

2. Author Gaurav Kumar declares that he has no conflict of interest.

3. Author Shrawan Kumar declares that she has no conflict of interest.

4. Author Bharti Takur declares that he has no conflict of interest.

5. Author Plakshi Saini declares that she has no conflict of interest.

6. Author Vanshika Mahajan declares that he has no conflict of interest.

References

1. Wikipedia. Línguagestualportuguesa. 2012 September 9, 2013 [cited 2013; Available from: http://pt.wikipedia.org/wiki/Lingua_gestual_portuguesa.
2. Vijay, P.K., et al., Recent Developments in Sign Language Recognition : A Review. International Journal on Advanced Computer Engineering and Communication Technology, 2012. 1(2): p. 21-26.
3. Mingqiang, Y., K. Idiyo, and R. Joseph, A Survey of Shape Feature Extraction Techniques. Pattern Recognition, 2008: p. 43-90.
4. Miner, R. RapidMiner : Report the Future. December 2011]; Available from: http://rapid-i.com/.
5. Mitra, S. and T. Acharya, Gesture recognition: A Survey, in IEEE Transactions on Systems, Man and Cybernetics2007, IEEE. p. 311- 324.
6. Murthy, G.R.S. and R.S. Jadon, A Review of Vision Based Hand Gestures Recognition. International Journal of Information Technology and Knowledge Management, 2009. 2(2): p. 405-410.
7. Wachs, J.P., H. Stern, and Y. Edan, Cluster Labeling and Parameter Estimation for the Automated Setup of a Hand-Gesture Recognition System. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2005. 35(6): p. 932-944.
8. Kumar, G., and V. Rishiwal. "Machine learning for prediction of malicious or spam users on social networks." *Int. J. Sci. Technol. Res* 9 (2020): 926-932.
9. Trigueiros, P., F. Ribeiro, and L.P. Reis. A Comparative Study of different image features for hand gesture machine learning. in 5th International Conference on Agents and Artificial Intelligence. 2013. Barcelona, Spain.
10. Trigueiros, P. and F. Ribeiro. Vision-based Hand WheelChair Control. in 12th International Conference on Autonomous Robot Systems and Competitions. 2012. Guimarães, Portugal.
11. Wang, C.-C. and K.-C. Wang. Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction. in Proceedings of the International Conference on Advanced Robotics 2008. Jeju, Korea.
12. Conseil, S., S. Bourenname, and L. Martin, Comparison of Fourier Descriptors and Hu Moments for Hand Posture Recognition, in 15th European Signal Processing Conference (EUSIPCO)2007: Poznan, Poland. p. 1960-1964.
13. Barczak, A.L.C., et al., Analysis of Feature Invariance and Discrimination for Hand Images: Fourier Descriptors versus Moment Invariants, in International Conference Image and Vision Computing2011: New Zeland.

14. Bourennane, S. and C. Fossati, Comparison of shape descriptors for hand posture recognition in video. Signal, Image and Video Processing, 2010. 6(1): p. 147-157.
15. Triesch, J. and C.v.d.Malsburg. Robust Classification of Hand Postures against Complex Backgrounds. in International Conference on Automatic Face and Gesture Recognition. 1996. Killington, Vermont, USA.
16. Boháček, Matyáš, and Marek Hrúz. "Sign pose-based transformer for word-level sign language recognition." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.
17. Xie, Pan, et al. "Multi-scale local-temporal similarity fusion for continuous sign language recognition." Pattern Recognition 136 (2023): 109233

Figures

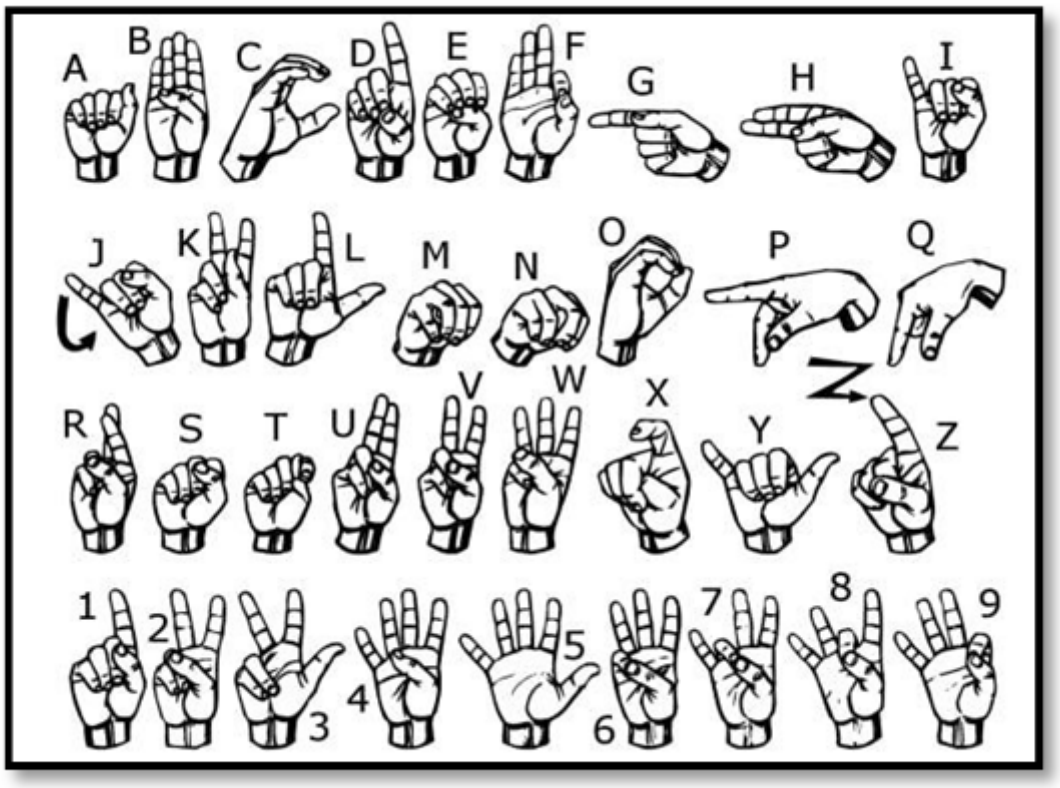

Figure 1

ASL Alphabets

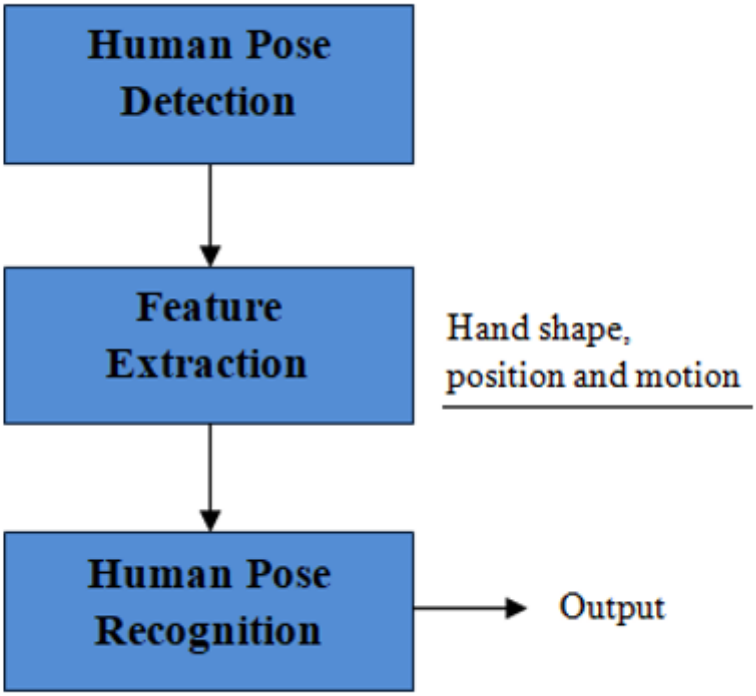

Figure 2

Functional Modules

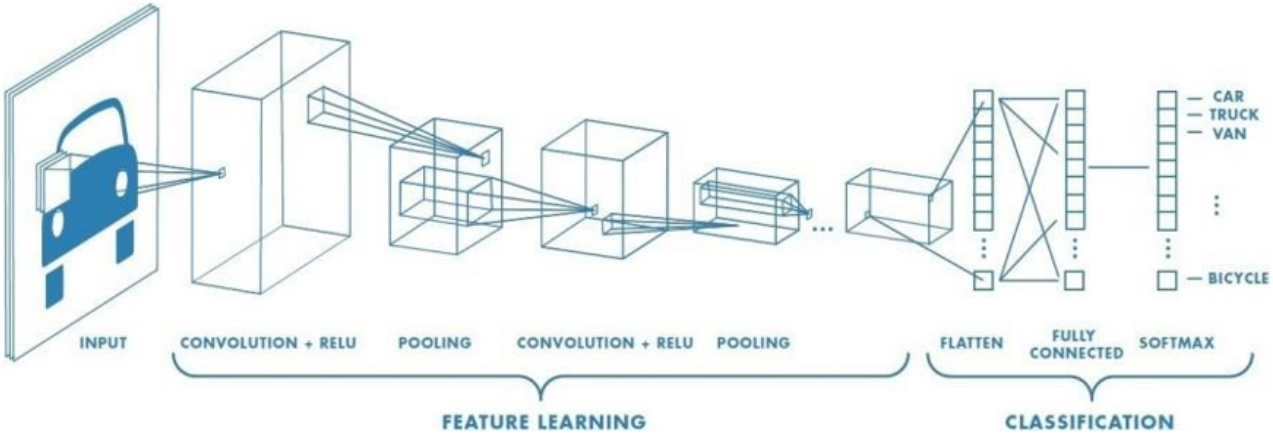

Figure 3

CNN

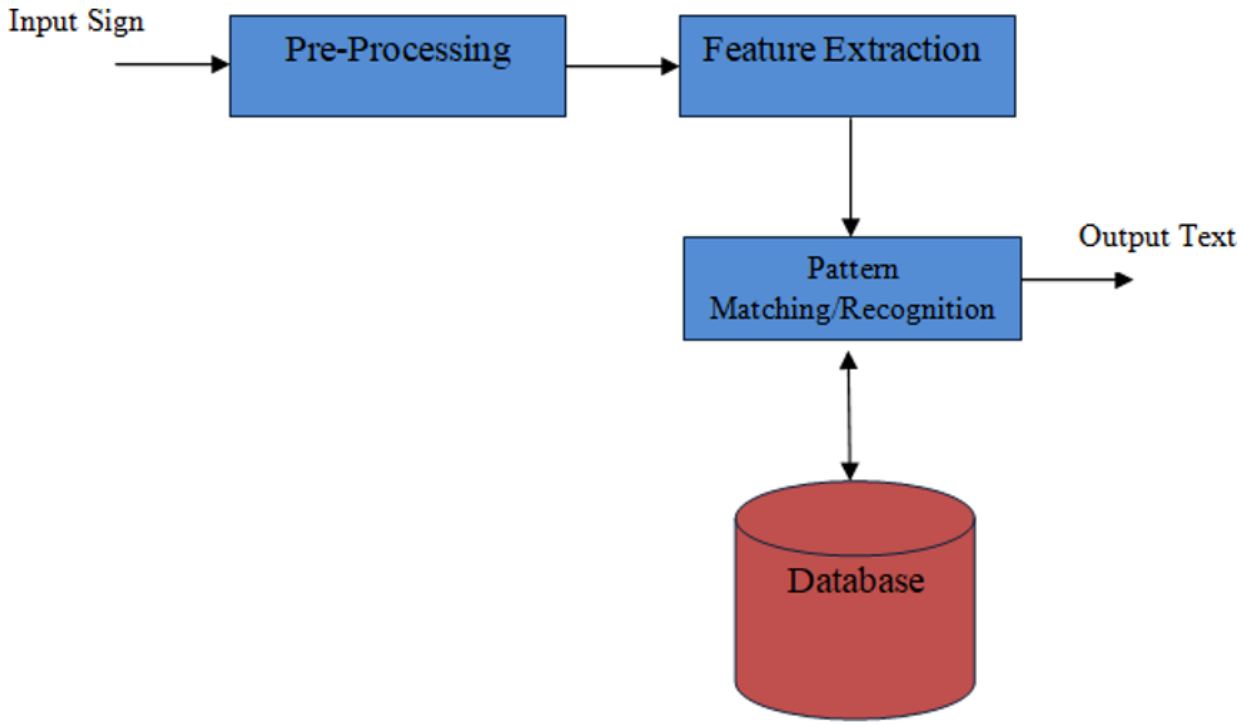

Figure 4

System Architecture

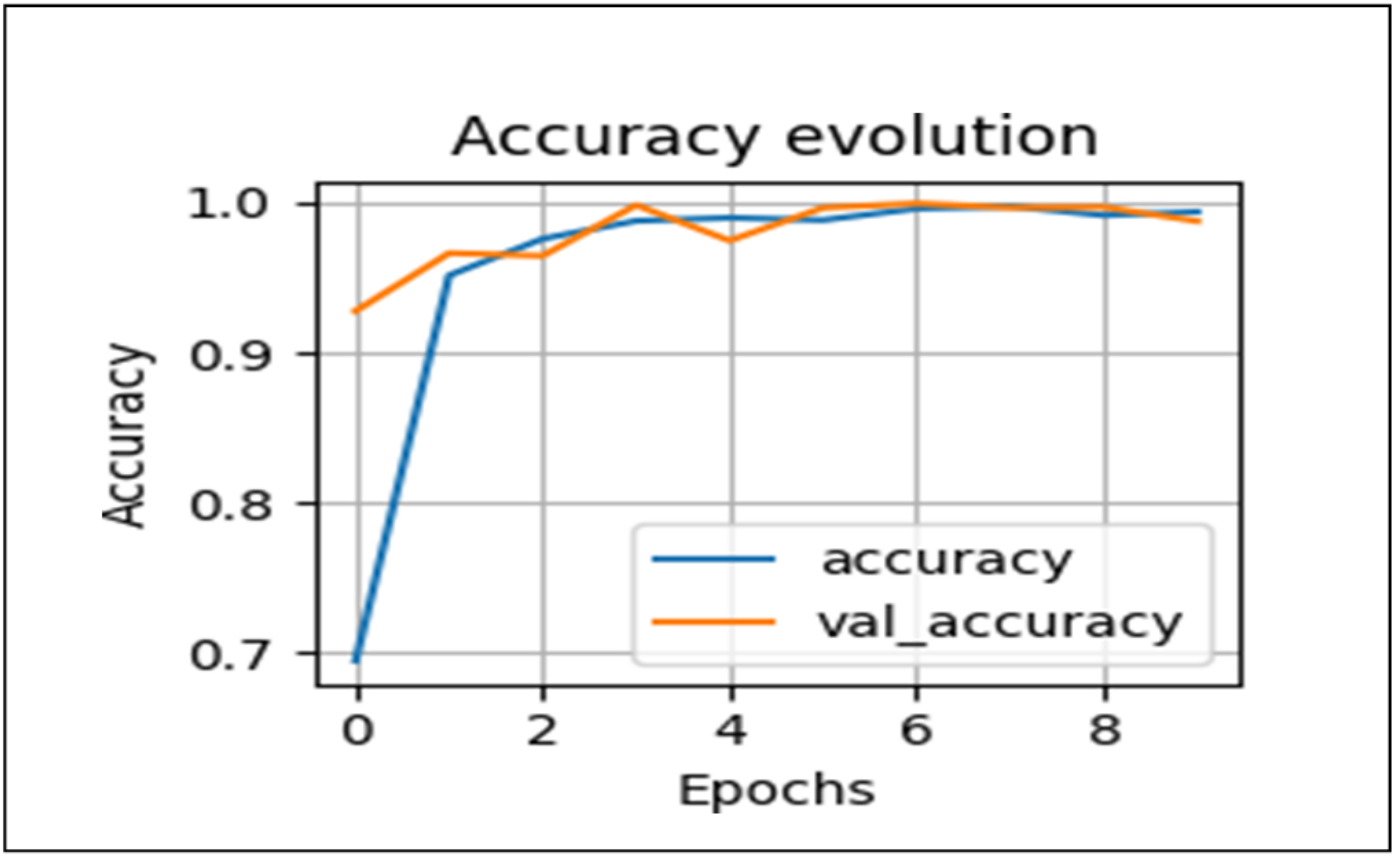

Figure 5

Plotting Training and Validation Accuracy vs. Epochs

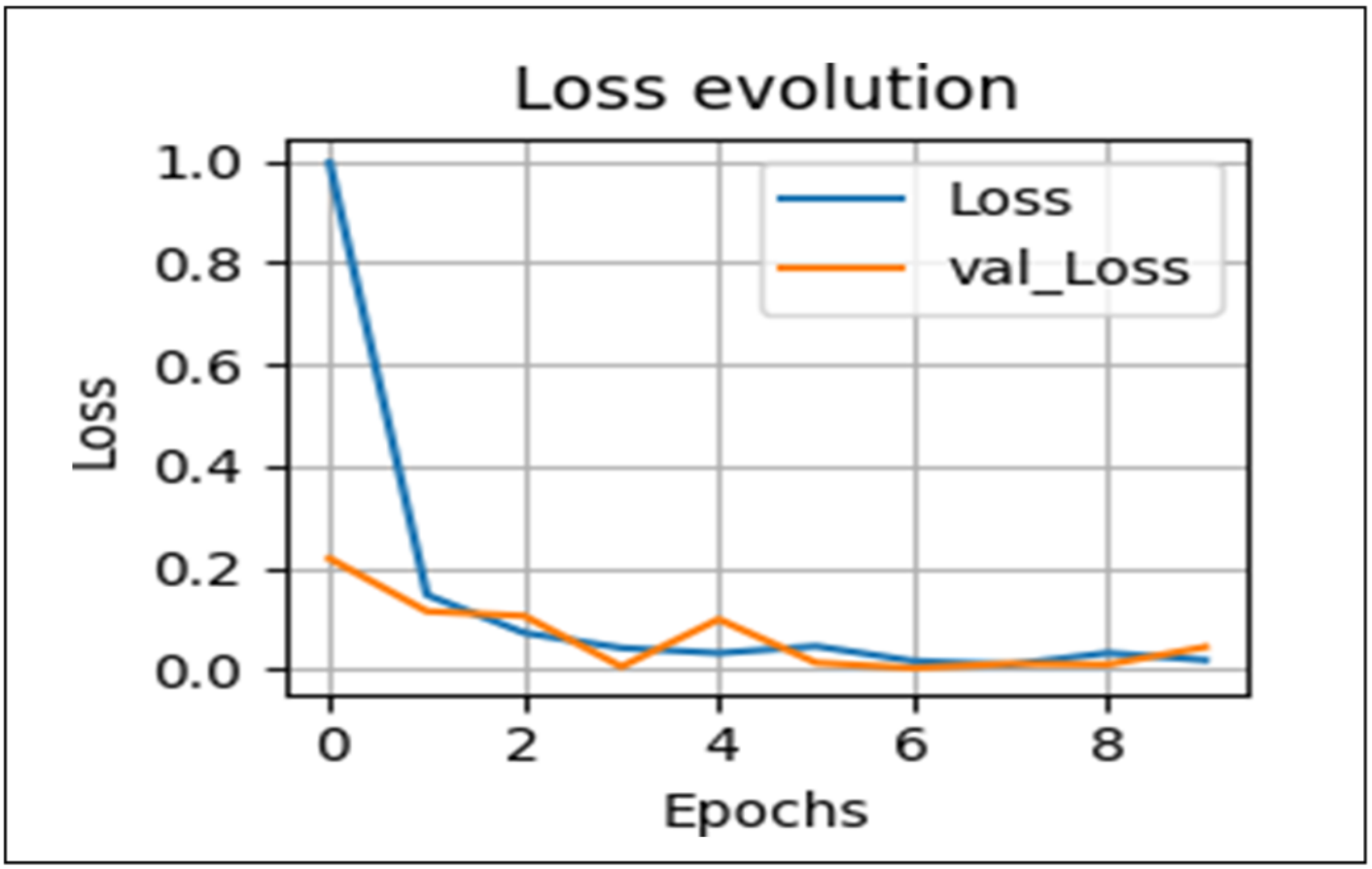

Figure 6

Plotting Training and Loss evolution vs. Epochs

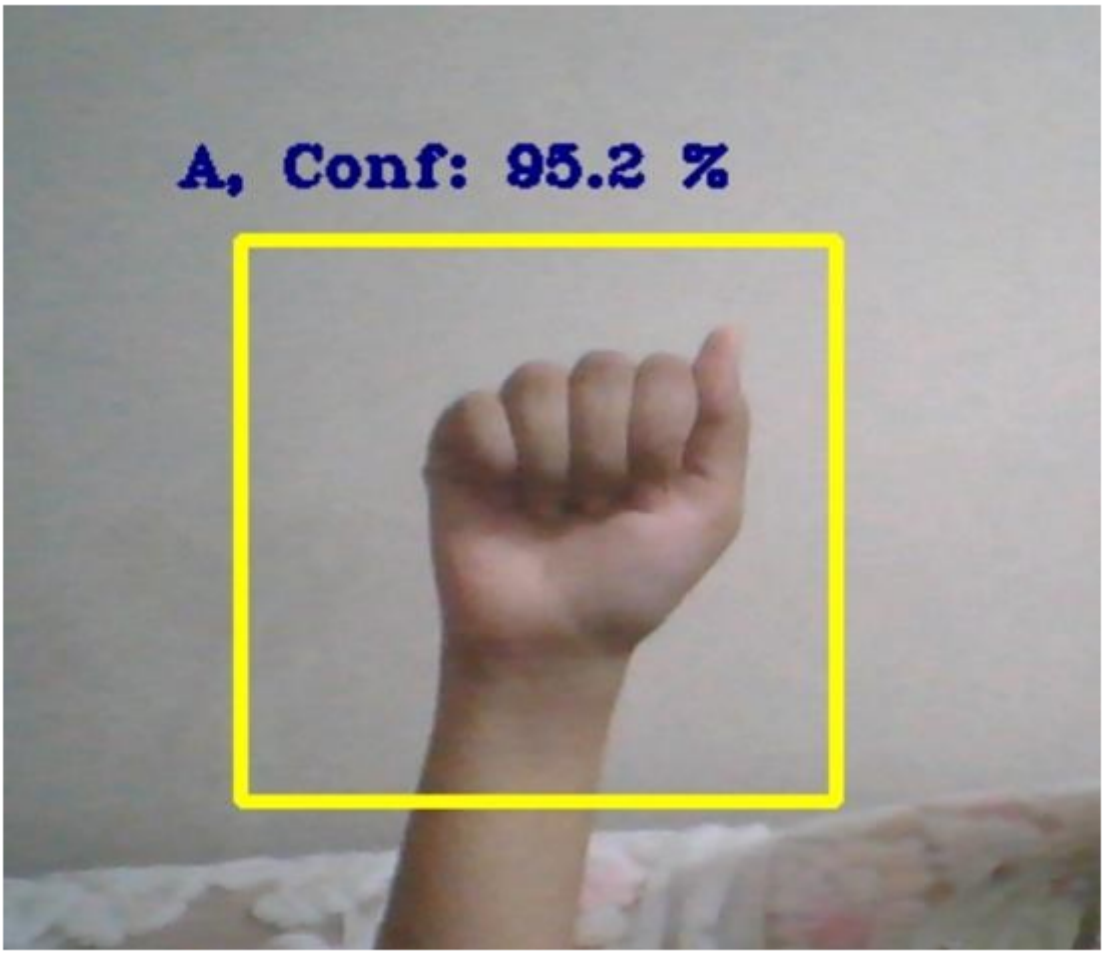

Figure 7

Letter A

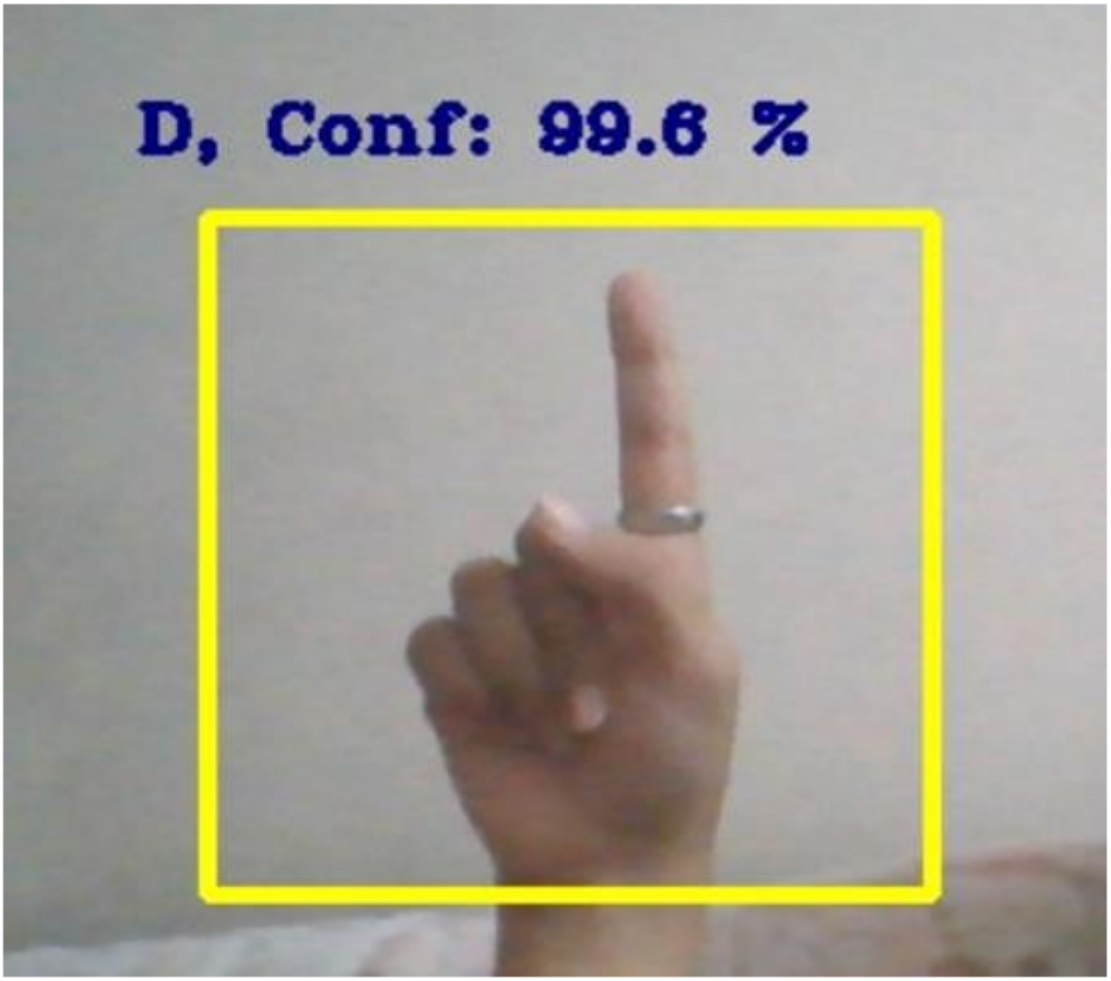

Figure 8

Letter D

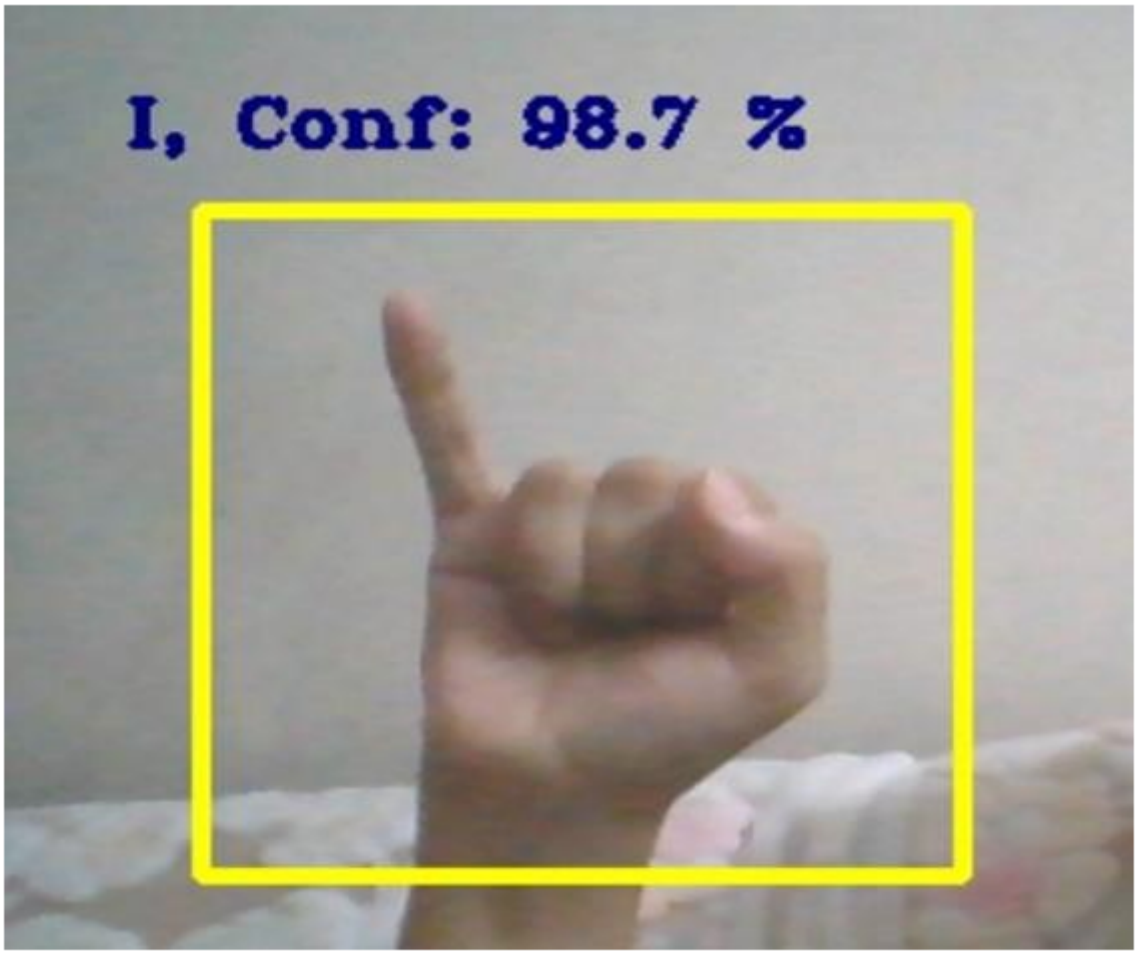

Figure 9

Letter I

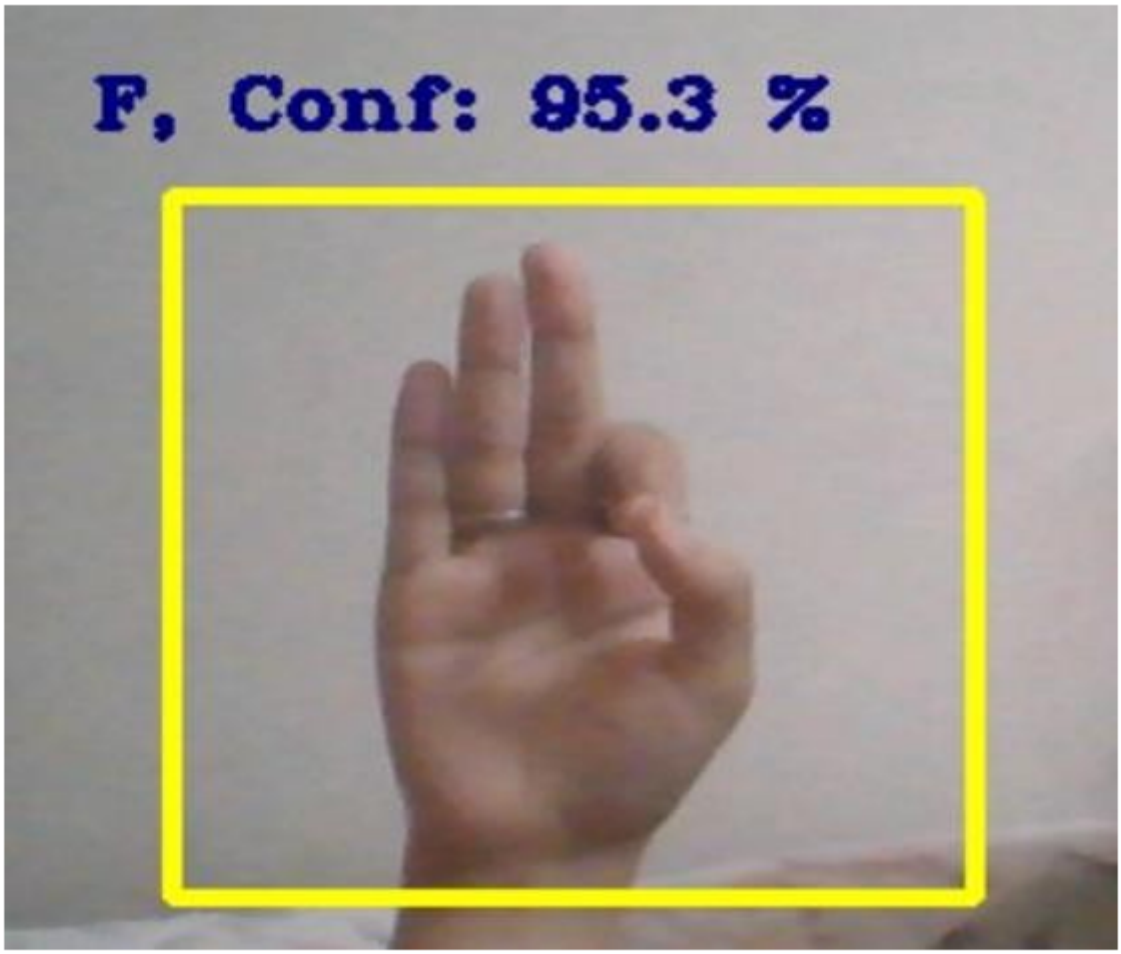

Figure 10

Letter F

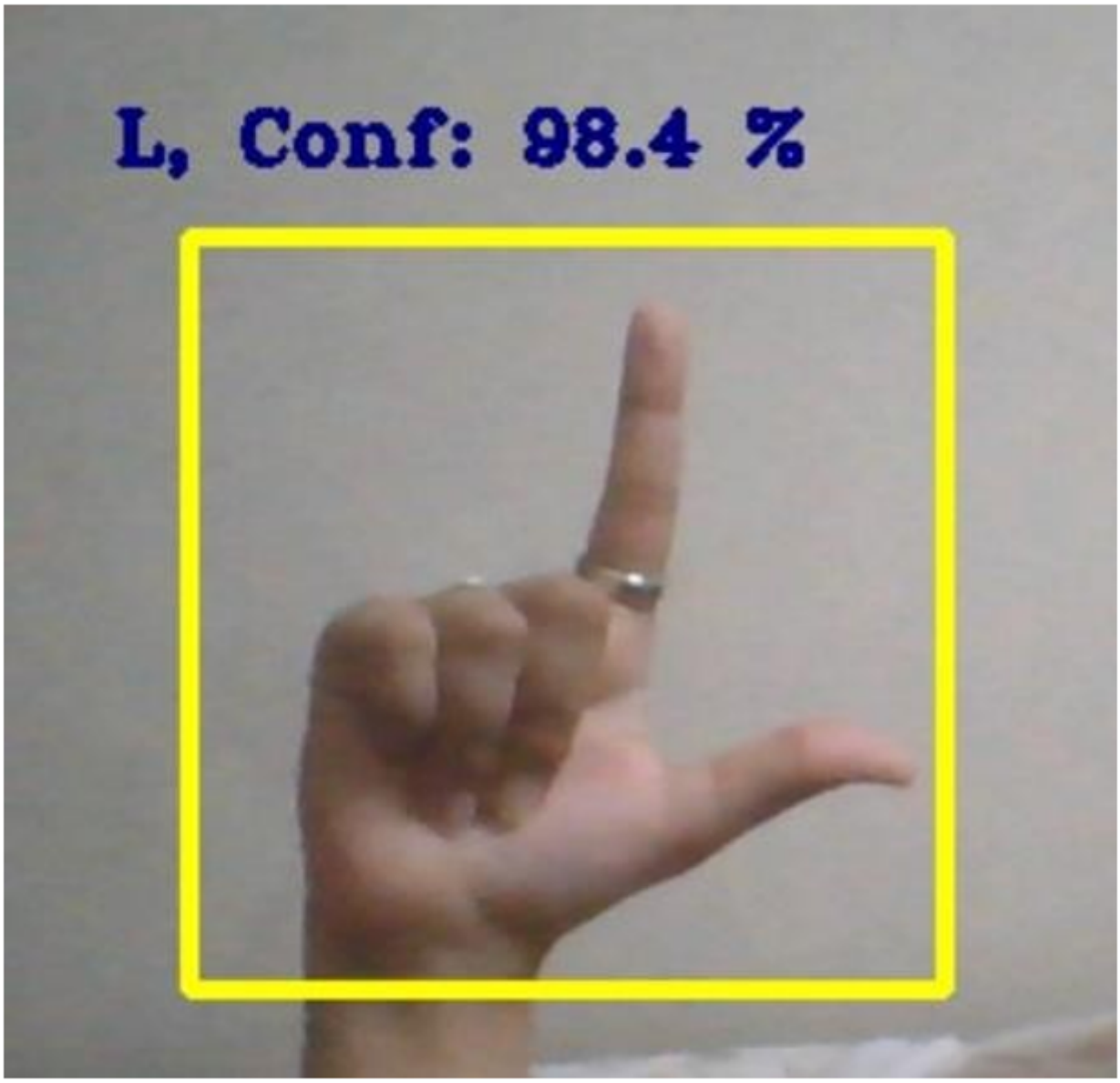

Figure 11

Letter L