Developing Amharic Sign Language Recognition Model for Amharic Characters Using Deep Learning Approach

Isayas Feyera (isayasfeyera@gmail.com)
Addis Ababa Science and Technology University https://orcid.org/0000-0002-4159-9609
Hussien Seid
Addis Ababa Science and Technology University

Research

Keywords: Amharic Sign Language, convolutional neural network, Faster R-CNN, Deep Learning, Single-shot detector, Convolutional Neural Network, Hearing-impaired people

Posted Date: April 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-236824/v1

RESEARCH

Developing Amharic Sign Language Recognition Model for Amharic Characters Using Deep Learning Approach

Isayas Feyera Olkeba* and Hussien Seid Worku

*Correspondence:
isayas.feyera@aastu.edu.et
Department of Software Engineering, Big Data and HPC CoE,Addis Ababa Science and Technology University, Addis Ababa, Ethiopia
Full list of author information is available at the end of the article

Abstract

Hearing-impaired people use Sign Language to communicate with each other as well as with other communities. Usually, they are unable to communicate with normal people. Most of the people without hearing disability do not understand the Sign Language and unable to understand hearing-impaired people. So, they need recognition of Sign Language to text. In this research, a model is optimized for the recognition of Amharic Sign Language to Amharic characters. A convolutional neural network model is trained on datasets gathered from a teacher of Amharic Sign Language. Frame extraction from Amharic Sign Language video, labeling and annotation, XML creation, generate TFrecord, and training models are major general steps followed for developing models to recognize Amharic Sign Language to characters. After training of the neural network is completed, the model is saved for recognition of Sign Language from a video system or from the frame of video. The accuracy of the model is the summation of confidence of individual alphabets correctly recognized divided by the number of alphabets presented for evaluation for Faster R-CNN and SSD. Hence, the mean average accuracy of the Faster R-CNN and Single-Shot Detector is found to be 98. 25% and 96 % respectively. The model is trained and evaluated for the character of the Amharic language. The research will continue to include the remaining words and sentence used in Amharic Sign Language to have a full-fledged Sign Language recognition model to a complete system.

Keywords: Amharic Sign Language; convolutional neural network; Faster R-CNN; Deep Learning; Single-shot detector; Convolutional Neural Network; Hearing-impaired people

Introduction

There are different ways of communication in the world in which people communicate with each other. Among these ways of communication, Sign Language is one of them. Sign Language is usually a means of communication among hearing-impaired people as well as between some normal and hearing-impaired people. Sometimes it can also be used among a few normal people. So, to make hearing-impaired people communicate freely with normal people, this Sign Language must be supported by technology to recognize the Sign Language or all normal peoples must be enforced to learn the Language. The latter option is much costly, if not impossible.

The Amharic language is the national language of the Ethiopian government which has been used as a written language for at least 500 years [1]. The Amharic language is spoken in Ethiopia with more than 17 million people. It is also spoken in other

countries with more than 400,000 people. Amharic is also spoken in Ethiopia with a very large family next to Afaan Oromoo, which is the largest language spoken in Ethiopia [1]. According to [1], the Amharic language is the most spoken language with a large number of speakers next to Arabic and Afaan Oromoo. Moreover, recognizing Amharic Sign Language to Amharic characters requires a high degree of technical skills in both Amharic and its Sign Language.

A computer can beat human beings better in mathematics and chess but less accurate in a visual perspective than a human being. An object can be identified in a complex scene with moderate success and even to find and name all of the people in a photograph can be attempted. But, because of the advanced nature of human beings, it is vivid that a computer might not be effective like a human being do.

Hearing-impaired seek basic needs like normal human beings do such as learning, teaching, reading, writing, communicating in which it may not be easy for them. Automatic Sign Language recognition is the technology that makes the computer to identify the sign used with the signer and convert to text with the help of some algorithm interfered with. Sign Language may also be the composition of mimic, gesture, hand sign, and fingerspelling in addition to the hand position.

A Greek philosopher called Socrates (469/470-399 BCE) once said that "If we had not a voice or a tongue, and wanted to express things to one another, would not we try to make signs by moving our hands, head, and the rest of our body, just as dumb people do at present?" This statement says that Sign Language is a natural language that the early community was using for their communication.

In the world, more than four hundred sixty-six million people who have lost their hearing as reported by world health organization [2]. It is over 5% of the world's population and among this number (5%) or 34 million of these people are children. Unless there is action to be taken with 2030 G.C., there will be about 630 million people with disabilities and hearing loss. The frustration in which this number will rise to 900 million by 2050 G.C also arises [3]. The problem of hearing-impaired or hearing loss indicates that the inability of hearing others' person speech either completely or partially. This problem may arise from a different source which may include nerve damage, extreme noise, disease such as a virus or even increasing the age of a person that may result in the problem [4]. Sign Language is a natural language that mostly hearing-impaired people use for communication. Researchers have tried to fill the gap of hearing-loss for the communities. A country such as Indian, Chines, Arab, and American has also tried to propose a better solution for hearing-impaired people. In our country, people who are hearing-impaired face different type of troubles which isolate him/her from social affairs. The outcome of isolation from social affairs results in loss of interest in education, loss of self-esteem, loss of self-confidence, the inability of communicating with surrounding, and the inability to communicate with their family too.

Literature Review

Sign Language is a natural language that mostly hearing-impaired people use for communication. Researchers have tried to fill the gap of hearing-loss for the communities. A country such as Indian, Chines, Arab, and American has also tried to propose a better solution for hearing-impaired people. In our country, people who

are hearing-impaired face different type of troubles which isolate him/her from social affairs. The outcome of isolation from social affairs results in loss of interest in education, loss of self-esteem, loss of self-confidence, the inability of communicating with surrounding, and the inability to communicate with their family too.

Researchers believe that Ethiopian Sign Language has originated from American Sign Language. When Ethiopian Sign Language was originated from American Sign Language, according to the SIL Ethiopia pilot survey (2005), out of 249 American Sign Language, twenty-five percent (25%) of Ethiopian Sign Language is originated from American Sign Language. This origination was having a point of view that the enhancement had been done based on the culture of Ethiopian people.

Machine Learning

Machine uses a powerful technique to learn from experience; this experience is built from a bunch of data [5]. This learning process experiences in the form of observational or interactions with the environment and performance learning improve. For instance, for Sign Language to be experienced by machine learning there should be a huge volume of data in the form of an image. Machine learning a subset of Artificial intelligence having many parameters of functions. These parameter functions are learned from data sets[6].

Deep learning

Deep learning is one of the powerful and among many popular methods for solving the problems of machine learning [5] [7]. Deep learning is deep in that it learns many layers of computation. For Sign Language, it is true as far as there is a deep layer, the accuracy of the model is becoming relevant. Consequently, there is a high probability of recognizing Sign Language to characters. The building block for recognizing Sign Language to characters arise from: the data and the pre-trained model which helps in transforming the data, a loss function that tells how much the model made mistake in predicting Sign Language frame [5].

Sign Language

The emergence of Sign Language in the 1960s in Ethiopia is the result of American and Nordic missionaries' appearances who opened the school of hearing-impaired [10]. There is also Sign Language comes from America and Nordic. Despite the miss conception that some people believe Sign Language is not a natural language, it is a natural language having its own rules. Sign Language can be used by hearing-impaired, dumb, or normal people like any other language. The importance of Sign Language arises from the fact that early human beings had been using this language before the advancement of vocal language along with computers today. A child before learning normal language use or express gesture communication whether it needs for food, warmth, or comfort. So, this Sign Language can also be learned at the very beginning with a gesture.

Sign Language is a vision-based language usually used by hearing-impaired people. The current situation exposes Sign Language is extensively used in international sports to read the plays Sign Language if used in an unwanted manner, in religious practice, for traffic board's on-road, and many other ways. Different country has

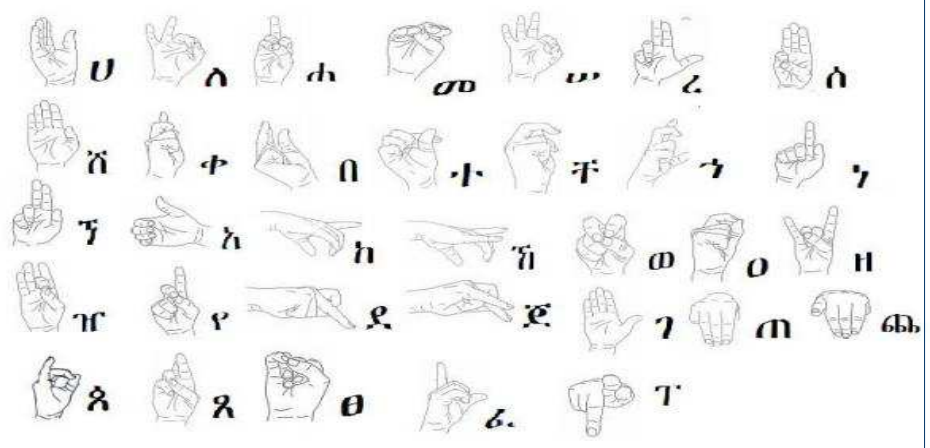

Figure 1 Amharic Sign Language Finger Spelling

different Sign Language structure: American Sign Language, Indian Sign Language, Arabic Sign Language, and more [8] [1]. In this 211th century, people need more comfortable and useful things to make their life easier. This is because of the field of science and technology. Even on the market, there are video games with real-time gesture recognition with the advancement of new technology provided that people who can play will be more advantageous. Although, not all people are fortunate but some peoples are physically challenged. This challenge could be deafness or being a phonic. So, researchers particularly computer scientists should help those people to help them communicate and express their filling with others.

Component of Sign Language

- Hand-shape
- Orientation
- Location
- Movement
- Facial Expression

Amharic Sign Language

Most Amharic Sign Language uses one-handed for all alphabets but anyone can use either of his or her hands [9]. Usually, finger spelling is used the Amharic Sign Language has no signs. The names of persons, countries, cities, and common words are signed by their first letter.The following picture indicates the finger spelling of each Amharic Sign Language.

Supervised learning

With supervised learning given the label of the Amharic Sign Language image as input, it predicts the target and recognizes its labeled Amharic characters. As its name indicates this learning type goes with supervision and provides the model with a data set containing labeled Sign Language where each sample matched with the

correct label. The interesting part is predicting the probability of Sign Language correctly classified. This justifies that a supervised learning guarantee for the majority of successful matching of Sign Language with a labeled image or frame [5]. Using the concept of supervised learning holds the feature of Sign Language image corresponding Sign Language characters. This value is called class which holds different Sign Language images. The usage of supervised learning for Amharic Sign Language is that the algorithm of supervised learning predicts the result based on the labeled datasets and can evaluate the accuracy of training data with little data that are not trained.

Unsupervised learning

In contrast to supervised, unsupervised will be difficult to apply for Sign Language. In the case of unsupervised learning, the model lets work a data inputted on its own to discover valuable information. Additionally, the input is given to the model without expecting the output desired . Since the image fed to the model is a frame extracted from the video of Amharic Sign Language, giving a huge amount of frame to the model and expecting recognized text is difficult. In other cases, why supervised learning is preferred Amharic Sign Language is that the data model uses labeled data whereas unsupervised learning uses unlabeled data.

Related Works

The study entitled "Hand Gesture Recognition Based on Single-Shot Multi-box Deep Learning" [10] has a scope of recognizing hand gestures to increase effective communication between humans and computers. The paper proposes a good approach to recognize hand gestures in a complex scene or indifferent environment using the Single-Shot Multi-box Detector algorithm. This algorithm has 19 layers and the data prepared is the benchmark of the database which is prepared by considering different background or for real-time hand gesture recognition system. Finally, the algorithm accuracy is found to be 99.2 % accuracy.

The Figure (a) above shows that the English character representation of hand gesture whilst the Figure (b) shows the final detection of character recognition with different background of the signer. The Figure (c) shows final recognition of the character "w" with good accuracy of 99.96%. The input size of image for the algorithm is 224 x 224 with RGB color. SSD is one of the state-of-the-art which detect a given object based on the feed-forward convolutional neural network instance found in those boxes. The paper uses three different background for recognizing English character gesture by collecting 1070 with for training. The model is also tested with 268 which is found in the training data sets. In testing phase characters "w", "o", "r", and "k" results good performance.

The method they used was able to achieve 98% accuracy for static hand poses. In paper of [30], the author designed a real-time human-computer interaction system based the Sign Language recognition called hand gesture. The proposed system consists of hand detection, gesture recognition, and human-computer interaction (HCI). They used CNN to recognize the gesture to identify complex gestures using only one camera. They have used another method called hand detection and background removal so that gesture recognition works well. At the level of gesture recognition,

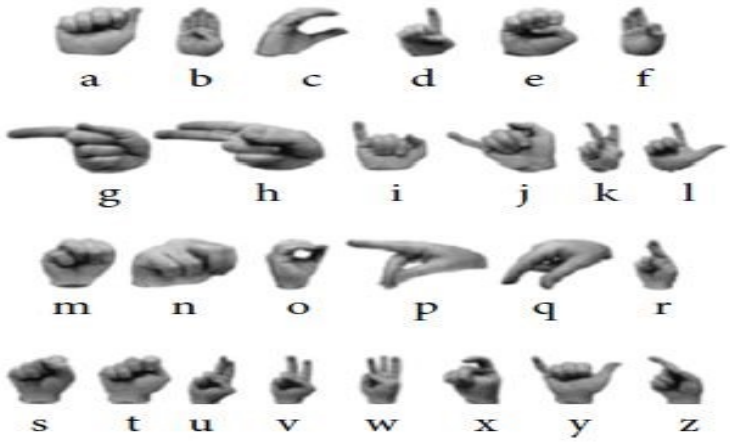

Figure 2 (a) English Finger Spelling

Figure 3 (b) Characters With Different Background

Figure 4 (c)Character Recognition

the CNN algorithm classifiers feed on the processed, binary images. Where the hand contour is centered and adjusted to a limited or fixed size. So, this algorithm trains a model using 16 classes of gestures. They collected a total gesture of 19,852 images from only five people. Each class has 1,200 sample images of Sign Language. The author used 200 samples for each as validation whilst 1000 are used for training the model achieving 99.8% accuracy. In the work [11] titled with Selfie video-based continuous Indian Sign Language recognition introduces real-time application for a mobile platform. The video is taken and processed by constraining its computing power for a mobile platform. To create a Sign Language feature space, they used pre-filter segmentation and feature extraction on the video frame. For feature space, they used two classifiers namely Artificial Neural Network and Minimum Distance Classifiers. These two classifiers are trained and tested on Sign Language feature space iteratively. There are three algorithms in which the authors use. The method called word-matching-score is applied to test the proposed Sign Language recognition and it results in the performance of the system 93.23%. The artificial neural network (ANN) performs the system with an accuracy of 90% in addition to a slightly small variation of 0.3 s in classification time compare to the above two methods. Hence, neural network classifiers with fast training algorithms and less classification time make the proposed method novel according to the author.

The paper published by Yigremachew Eshetu [3] discuss about a real-time Ethiopian Sign Language translation to audio converter with a hybrid of vision-based and sensor-based area unit to capture hand configurations and knowledge of the corresponding meaning of gestures. The paper uses the pointed method for the following purpose discussed separately. The first one is a vision-based approach: they use this with a single camera to track the user's hands for recognizing Ethiopian Sign Language The system uses a machine learning neural network called Single Shot Multi-Box Detector (SSD) on TensorFlow. This network helped them to detect the hand gesture. The main part of the steps for this vision-based sensor is face detection by using haar-cascade [12] [3] to justify whether there are who are going to sign, hand detection, and overlap detection. This overlap detection indicates that whenever the system takes a video from real-time it identifies and analyzes the video if there is an overlap of hands and face in Ethiopian Sign Language meaning. For example, if someone puts his hand on his chin this is to mean in Ethiopian Sign Language 'Mother' and if some put his/her hands on his/her forehead this is to mean 'Father' as shown in the following pictures respectively. This indicates that according to the method of SSD, hands, and forehead are overlapped so that there is Amharic Sign Language meaning depending on the place of hands the user located.

Methodology

This section discusses the procedure or techniques used to identify, select, and process as well as analyze data about Amharic Sign Language and its recognition of characters are discussed. The following methodology is used to collect data in the form of video, preprocess, and feed the prepared data for our model is also explained.

Data collection

Relevant Sign Language image and video which is needed to conduct the research has been collected. The dataset is useful for training and evaluating the system such

Figure 5 Detection overlap example (it means 'Father') [3]

that the recorded video by a mobile camera in a typical the same environment. The collected data by recording the teacher of Sign Language is prepared by extracting its frame.

Data set preparation

Signing Amharic Sign Language is a challenging task; since it needs professionalism or experience. The recorded video of the Sign Language with a mobile camera each alphabet separately was led to wait for 20 seconds to add the orientation of the alphabets. From the video collected, the frame of a video is extracted. Each 10 Amharic Sign Language characters has its class. In each Sign Language class, there are 500 frames of images, from this number, 80% of the total image for training the model is categorized and 20% of the total image for testing the model is categorized.

Video Processing

The preprocess plays a crucial role in deep learning in getting a better result of the model. Preprocessing included resizing the whole images to 244, 244, 3. On the resized image the number 3 indicates or represents RGB or (Red, Green, and Blue) and it is called the 3 channels. The next task is splitting the images to train and test. Machine learning API called KERAS on VGG-16 pre-trained model is used. The data is labeled and annotated by using the labelImg tool. Five hundred (500) hundred images for each class are labeled. After labeling and annotating the images, this tool converts the labeled image to the XML form. The conversion of XML to CSV or to comma-separated file should also be done in which the data is ready to be trained.

Result and Discussions

For this study two different experiments are done by regularizing and fine tuning the neural network. For Regularization of the neural network, dropout regularization used. It is used for reducing over fitting of the mode by adding the value of 0.5.The addition 0.5 dropout helps in random random removal of 50% of the units from each hidden layer and finally end up with as simpler network.

Figure 6 Extracted Frame Annotation Using LabelImg tool

Hyper-parameter for Fine-tuning Model

In any machine learning algorithm, there is a high difference between model hyper-parameters and model parameters [13] . Model training is insisted on by the machine learning model or by the classifier to learn on its own. This case is called the model parameter. Whereas model hyper-parameter controls the process of training. The model parameter is set before feeding the data to the model or before starting the training of the data. The behavior of the algorithm can be directly controlled by hyper-parameter since the performance of the model under training can be affected by it.

Learning rate

Among hyper-parameter available for deep learning, the learning rate is the most important hyper-parameter. If the learning rate set is too small, it will take a much longer time (more than a thousand) of epochs to reach an ideal state. In other cases, if the learning rate set is larger value then the algorithm may not converge. By considering these two conditions learning is initialized rate to be 0.001.

Number of epochs

Different number of epochs is done for recognition of Amharic Sign Language to text. But to select the right number of epochs from the training stage, validation error should be considered. One can also use another technique called early stopping in case of validation error has not improved after some specific epochs. We used 100 number of epochs for recognition of sign language to text.

Batch Size

In neural network there is a difference between batch size and number of batches. Batch size is total training image which found in a single batch. It is difficult to pass

Table 1 Bounding Box and Confidence of Amharic Sign Language

Model	Batch Size	Epochs	Iterations	Learning Rate	Optimizer
Faster-RCNN	16	-	49809	0.0002	Momentum
SSD	16	-	46051	0.0002	Momentum

the whole sig language to the model at once; the data should be broken down a given number of batches. In this research by considering the machine batch size is set to 16. This 16 indicates the number of sign language pass to the neural network among the whole sign language prepare for training. It is also possible to get number of epochs from iterations the neural network undergone. Iteration indicates how much number of batches are required to finish one epoch. In neural network, one epoch refers when the whole input image fed to the neural network through forward and backward only once. The momentum value we used is a simple technique used in neural network to increase training speed and accuracy. In neural network the aim of the training resides in the fact to find the values for weight and bias provided that for set of input data there is predicted match output.

Amharic Sign Language Character Recognition Using Faster R-CNN

the hypothesizing of object location, there is a need of state-of-the-art which depends on the region proposal such as Faster R- CNN and SSD [14]. There are so many state-of-arts for object detection such as R-CNN, Fast R-CNN, YOLO (You Look Only Once). Each of them differs from each other in either speed or accuracy of detecting an object. For example, the algorithm R-CNN consumes an extreme amount of time training than the rest of the state. This is because 2000 Region Proposal Network (RPN) per image should classify and it will take 47 seconds for each test image [15] in real-time implementation. To solve this problem, a Fast R-CNN object detection algorithm was developed. The main difference between Fast R-CNN and R-CNN is that, in the case of R-CNN, the region proposal feeds to CNN but in the case of Fast R-CNN the input image is fed to CNN to generate a convolutional feature map. Then from this convolutional feature map, the region of the proposal should be identified to squares. Then by using the RoI (Region of interest) pooling layer, the region proposal should be reshaped to a specific size to feed it into a fully connected layer. Finally, by using the Softmax layer, the model predicts the class of the proposed region and also offset values for the bounding box set during annotation and labeling of Amharic Sign Language.

The reason Fast R-CNN is better in speed than R-CNN is that 2000 region proposals are fed to CNN every time rather the convolution computation or operation will be done only once per image, and a feature map is generated from it [15]. There is also the YOLO algorithm which faster than any of the three algorithms (R-CNN, Fast R-CNN, and Faster R-CNN) . The research we are doing do not consider YOLO for Amharic Sign Language recognition. The reason is that the study does not only consider speed rather the accuracy of detecting Sign Language is a critical matter since Amharic Sign Language is similar and it is extremely difficult for the neural network to identify one Sign Language from other. YOLO is more accurate but struggles with a little object found on the image. And it should also consider the whole image but Faster R-CNN only considers the RoI annotated during labeling of the image.

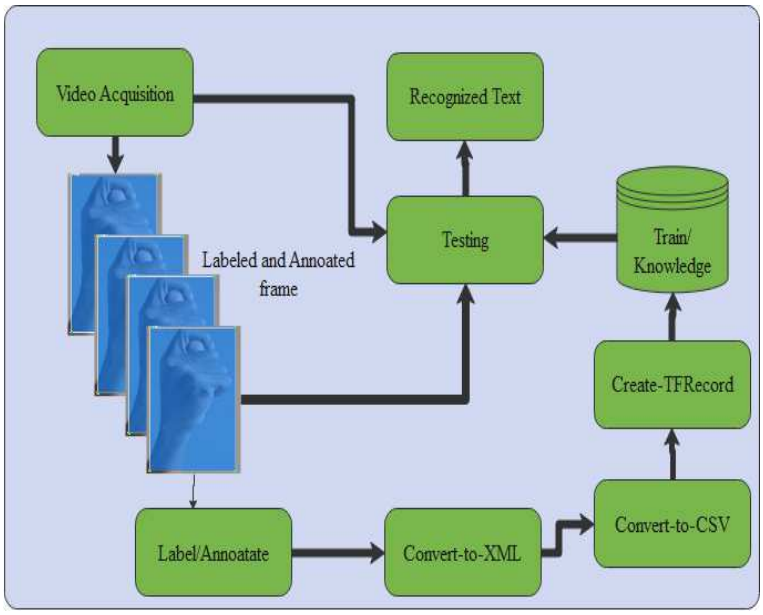

Figure 7 Data annotation and labeling

Data Labeling, Annotation and Training

Data annotating is one of the most important steps in making the portion of Sign Language understandable and recognizable to the model. Machine learning and artificial intelligence seek this step to learn patterns and memorize the same for good prediction. Data annotation or data labeling is making a selection for some portion of the image to be recognized by the algorithm. Annotation aims to make Sign Language recognizable to visual perception for the deep learning algorithm selected. General step for frame extraction for annotation After a recording of Sign Language video from a signer of Amharic Sign Language, the frame of each movement of the signer sequentially is taken from video. After taking each sequence of a frame from the video the frames are annotated as much as possible for model training.

Evaluating Faster R-CNN

In this study, for evaluating the optimized Faster R-CNN model, mean average precision (mAP) is used. It means nothing than the average of precision seen in the model for recognizing correct Amharic sign language characters. In computer vision mAP is used for localizing and classifying the object to identify the where the object is located or the bounding box coordinates also with what type of the object it is.

True Positive (TP): indicate the sign language predicted is correct or the sign language is predicted as positive as was correct.

False Positive (FP): indicates that the sign language predicted is positive but it is incorrect.

Intersection over union (IoU): to measure the how much two boundaries of

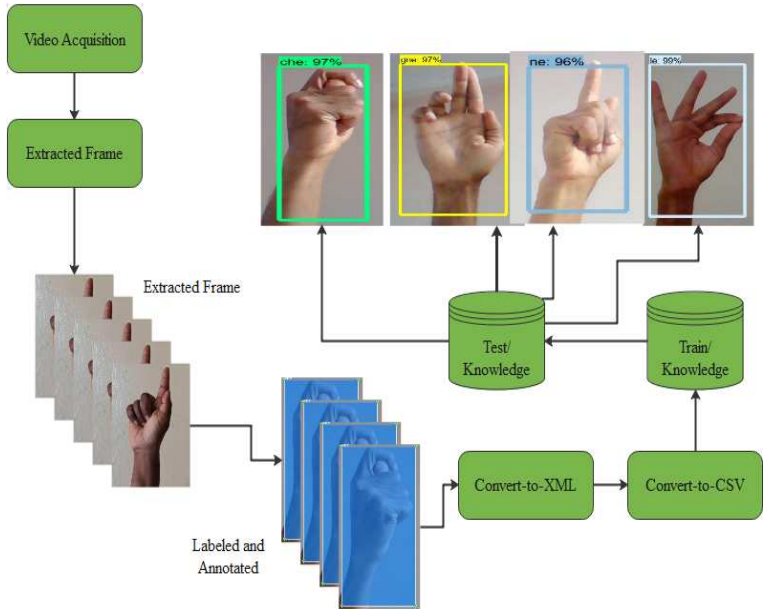

Figure 8 Recognized Character Using Faster-RCNN

sign language overlap with the ground truth or the real sign language boundary. Since the sign language region of interest is annotated while labelling the region of interest. To classify whether the predicted sign language is true or false 0.6 threshold value is used.

$$Intersectionoverunion(IoU) = \frac{areaofoverlap}{areaofunion} \tag{1}$$

Loss Function for Faster-RCNN

Unlike SSD, Faster R-CNN has improvements such as multi-scale feature extraction and default boxes. Faster R-CNN is optimized for multi-task loss function like another state-of-the-art do such as Fast R-CNN. The following equation clarifies how the Faster R-CNN got the loss function which is also shown on the following figure The loss classification and bounding box are combined with multi-task loss Other important progress can be seen on the following graph the loss graph which shows the overall classifier of Sign Language with increasing time. The Y-axis indicates that the loss value which is significantly decreasing provided that the accuracy of the model getting better and better. Whilst the X-axis indicates that the number of iterations undergoes for training the model.

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{box}$$
$$\mathcal{L} = \mathcal{L}_{cls} \mid + \mathcal{L}_{box}\mathcal{L}\left(p_i, t_i\right) = \frac{1}{N}\sum_i \mathcal{L}_{cls}\left(p_i, p_i^*\right) + \frac{\lambda}{N_{box}}\sum_i p_i^* \cdot L_1^{smooth}\left(t_i, t_i^*\right)$$

Where $\mathcal{L}_{\text{cls}}$ indicate that the logarithm of function over two classes. This is true in that it is possible to translate multi-classification into a binary classification by

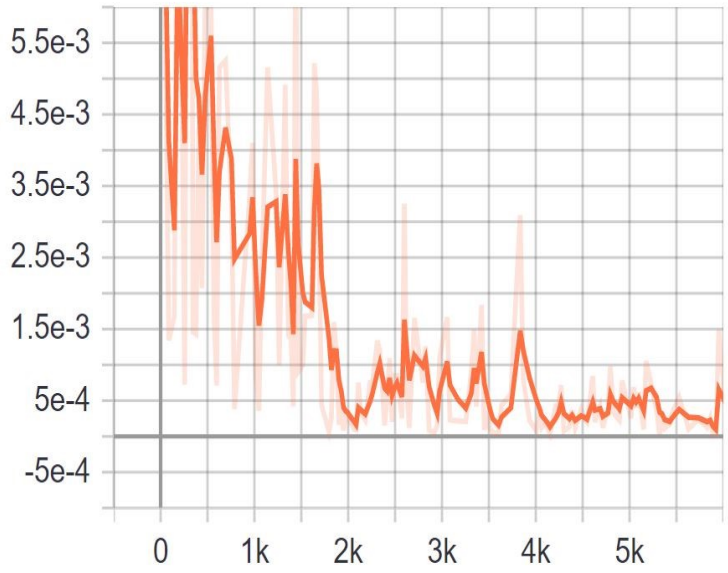

Figure 9 Loss Function of Faster-RCNN

predicting the object target is correct or not. L_1^{smooth} is the smooth L1 loss.

$$\mathcal{L}_{cls}\left(p_i, p_i^*\right) = -p_i^* \log p_i - \left(1 - p_i^*\right) \log\left(1 - p_i\right)$$

Amharic Sign Language Recognition using Single-Shot Detector (SSD)

Using a neural network for the case of Sign Language recognition will enable prediction of the sign in an image. But hand recognition not only predicts the hand but also a probability of there is a hand based on an annotated image of a hand, the height of the bounding box, the width of the bounding box, horizontal coordinate of the center point of the bounding of box, and vertical coordinate of the center point of the bounding box. SSD has two component backbone models and the SSD head is a pre-trained model as a feature extractor [16]. Unlike SSD, Faster R-CNN runs the whole process at 7 frames per second (PFS). SSD adds some improvement than Faster R-CNN such as multi-scale feature extraction and default boxes. The reason SSD speed higher than Faster R-CNN came from the fact that the improvement of multi-scale feature extraction and default boxes using lower resolution images. The accuracy of SSD is measured by mean average precision (mAP). SSD is detection is folded into two types: extract feature maps and apply convolutional filters to detect objects [16]. SSD uses visual geometry group16 (VGG-16) for extracting the feature map of Sign Language we used. Following the extraction of features from the image it detects the object using a convolutional layer. Unlike other state-of-the-art such as Faster RCNN, SSD does not use region proposal network (RPN) rather by using small convolutional filters it computes the location and class scores. SSD has a similar feature with CNN SSD apply 3 x3 convolutional filters to each cell for

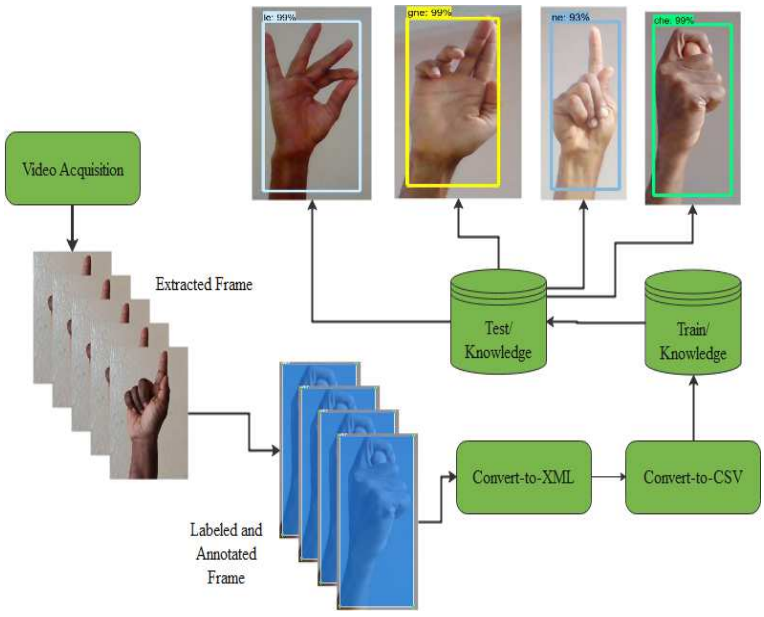

Figure 10 Recognized Character SSD

predicting the object [48].

Training Phase

There is a difference between SSD and other detector using region proposal in that SSD use ground-truth information that needs to assigned as specific output in the fixed set of detector outputs liu2016ssd. The training objective of Single-shot Detector is derived from the Multibox object detector but extended to handle multi-box object categories. SSD has two options for using an object and detect based bounding box trained on. These are 300 x 300 or 512 x 512. Considering the computational machine for this study 300 x 300 is used. For training SSD, it takes about 96 consecutive hours or around 4 consecutive days with Forty-six thousand Fifty-one (46051) iterations are done.

Loss Function for SSD

Whenever a neural network is trained or being training, there is an error that occurred incorrectly classifying the object. This emphasizes whenever the entire datasets are passed forward and backward via a given neural network the output may result in an error [50]. This is called localization loss. SSD ignores the negative match of the Sign Language it detects only penalizes from positive matches. The localization loss between the predicted box l and the ground truth box g is defined as the smooth L1 loss with cx, cy as the offset to the default bounding box d of width w and height h.

Table 2 Bounding Box and Confidence of Amharic Sign Language

Model	Frame	Detection	Confidence	TP or FP
SSD	le_Frame	le	99%	TP
	gne_Frame	gne	99%	TP
	ne_Frame	ne	93%	TP
	che_Frame	che	99%	TP
	qe_Frame	qe	94%	TP
Faster-RCNN	le_Frame	le	99%	TP
	gne_Frame	gne	97%	TP
	ne_Frame	ne	96%	TP
	che_Frame	che	97%	TP
	qe_Frame	qe	95%	TP

Discussion

The main challenge in this thesis was data collection and the labeling of data collected. And also pickling or the one-hot making of data took more than 14 for the case of VGG-16 and more than 24 for the case of Faster R-CNN and SSD. This due to a large number of frames extracted from the video captured. In this work, three experiments were undergone as part of the study to evaluate Sign Language recognition. The first was done by using the VGG-16 pre-trained model and the second experiment was done with the help of Faster R-CNN and SSD. The experiment done on the VGG-16 model show a better result than the CNN model with equal and different amounts of datasets. For the recognition of Amharic Sign Language with VGG-16, three general steps are followed. The first step is labeling the class of the Sign Language followed by the one-hot encoding of the Sign Language with the respective class. Finally, train the model and load for the recognition of the Sign Language. For recognition of Amharic Sign Language to text with the Faster R-CNN and SSD algorithm, four general steps are followed. The extraction of a frame from video, labeling, and annotating and XML conversion of the annotated image followed by training and testing of the model will conclude the step.

Bounding Box

The above Table shows the bounding box of Amharic sign language along with its corresponding percentage of the confidence. The last column illustrates whether the detected sign language is True Positive or the detected sign language is False Positive. For this thesis, True Positive is set be considered if the intersection of the union or (IoU) is greater than or equal to 75% otherwise it is considered as True Negative. ***Precision:*** is used for justifying how much percentage the sign language recognized is predicted. For testing of sign language there are a lot of steps should be followed. One of these steps is finding precision of each characters. Precision is can found by dividing True positive of each sign language by the summation of True Positive and False Positive.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall: Recall of the characters can be found dividing the number of True Positive with the summation of True Positive and False Negative.

$$Recall = \frac{TP}{TP + FP} \tag{3}$$

Mean Average Precision (mAP): To evaluate the performance of the detector or the performance of the model, mean average Precision if taken into account. The reason why average precision included is that normally Average Precision only used for one class. But the sign language recognized in this research is more than one class the mean of each class recognized should be calculated by using the following formula.

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{4}$$

where K=10, number of Sign Language class to be recognized and AP is average precision. Hence, mAP can be calculated as mean of the average precision of across all 10 class of sign language.

Conclusions

The purpose of this study will be concluded by the recognition of Amharic Sign Language to Amharic character with two optimized algorithms namely the Faster R-CNN and SSD algorithm. Faster R-CNN, and SSD are used by adopting the Amharic recognition with equal size of data sets to identify which model is better in terms of speed and accuracy. Hence, in this study, it is realized that Faster R-CNN is better in accuracy for recognizing Amharic Sign Language. In other cases, SSD is better in speed compared to Faster R-CNN but less accurate in recognizing Amharic Sign Language. The model has been tested using the frame/image and video of Amharic Sign Language Characters. Faster R-CNN model and SSD was able to detect and recognize the Sign Language resulting with test different accuracy which is 98. 25 % and 96 %

Recommendations

The study shows that Sign Language recognition is a wide area that has been tried by many researchers. Since research is a matter of re-searching, the model developed will fill some gap that has not been done yet. Particularly coming to Amharic Sign Language, there is plenty of gaps that need to be done in the future including real-time classification of Amharic Sign Language recognition to texts. By considering the time and resources the study does not include the following important concept. So, at the moment the following concept are recommended for further study:

- The static background is used while capturing a video of Sign Language signer. So, it is recommended for the future to use dynamic background video.
- Researchers should also take into consideration using the whole Amharic Sign Language characters to fully recognize Amharic Sign Language. item The use of word and sentence level of Amharic Sign Language is also highly recommended.
- Recognition of Amharic Sign Language which has the motion to be classified and/or recognized with real-time to characters is also recommended.
- A huge number of videos of hearing-impaired and normal people should be used with different backgrounds for better results of Sign Language recognition to characters.

Table 3 Abbreviations and Acronyms

Abbreviations and Acronyms	
ASL	American Sign Language
CNN	Convolutional Neural Network
EthSL	Ethiopian Sign Language
Faster R-CNN	Faster Region-Based Convolutional Neural Network
SSD	Single Shot Multi-box detector

Acknowledgments
First and foremost, I would like to thank almighty God for his strongest love and help. I would also like to forward my deepest respect and appreciation for my advisor Dr. Hussein Seid for his ultimate help and guidance to this work. This thesis gets to end with the help of a Sign Language teacher found at Adama Elementary school Teacher Ayinalem and teacher Ayelech who are Sign Language teacher at Sebeta Voluntary school and also a reporter of Oromia Broadcasting Network Television (OBN). Finally, I would like to thank all my family, friends, and colleagues for their motivation and push me to go through this study.
. . .

Funding
The journal of big data is funding for African country to publish the research.

Abbreviations
Some of The agronomy we used is listed below. . .

Availability of data and materials
Data collected for this study is published on Mendeley data available at http://dx.doi.org/10.17632/5d3nkyhsrf.1. . .

Ethics approval and consent to participate
The text written and acknowledged is true and is the fact we have got by experiment.. . .

consent to participate
The Sign language image and video we gathered is with permission of the concerned part.As we can see from Figure 5 and 6 on this manuscript of page 8 and 9. Hence we get the permission from them about the consent of publication.

Competing interests
We confirm that there are no competing interests. The article is not under any other review process and is not in the subject of any other submission.

Authors' contributions
Contribution to the knowledge. This study has the following contributions:

- Preparation of corpus or data sets for Amharic Sign Language.
- Development of three different models for Amharic Sign Language.
- Use of pre-trained model with the concept of transfer learning for Amharic Sign Language.
- Representing image with text called Sign Language recognition to Amharic characters.

Authors' information
Text for this section. . .

Author details
Department of Software Engineering, Big Data and HPC CoE,Addis Ababa Science and Technology University, Addis Ababa, Ethiopia.

References

1. D. F. WOLDE, "Machine translation system for amharic text to ethiopian sign," Ph.D. dissertation, Addis Ababa University, 2011.
2. S. Chadha and A. Cieza, "World health organization and its initiative for ear and hearing care," *Otolaryngologic Clinics of North America*, vol. 51, no. 3, pp. 535–542, 2018.
3. Y. Eshetu and E. Wolde, "A real-time ethiopian sign language to audio converter."
4. R. Jiang, Q. Lin, and S. Qu, "Let blind people see: real-time visual recognition with results converted to 3d audio," *Report No. 218, Standord University, Stanford, USA*, 2016.
5. A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," *Unpublished Draft. Retrieved*, vol. 19, p. 2019, 2019.
6. D. Gunning and D. W. Aha, "Darpa's explainable artificial intelligence program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.
7. L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu *et al.*, "Deep learning for generic object detection," *A Survey [J].*, 2018.
8. M. Tesfaye, "Machine translation approach to translate amharic text to ethiopian sign language," 2010.
9. A. Bauer, "Eyasu hailu tamene, the sociolinguistics of ethiopian sign language: A study of language use and attitude. washington, dc: Gallaudet university press, 2018. pp. 160. hb. 60," *Language in Society*, vol. 48, no. 2, pp. 313–314, 2019.
10. P. Liu, X. Li, H. Cui, S. Li, and Y. Yuan, "Hand gesture recognition based on single-shot multibox detector deep learning," *Mobile Information Systems*, vol. 2019, 2019.

11. G. A. Rao and P. Kishore, "Selfie video based continuous indian sign language recognition system," *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 1929–1939, 2018.
12. M. B. B. Frew, "Audio-visual speech recognition using lip movement for amharic language."
13. X. Xiao, M. Yan, S. Basodi, C. Ji, and Y. Pan, "Efficient hyperparameter optimization in deep learning using a variable length genetic algorithm," *arXiv preprint arXiv:2006.12703*, 2020.
14. K.-H. Shih, C.-T. Chiu, J.-A. Lin, and Y.-Y. Bu, "Real-time object detection with reduced region proposal network via multi-feature concatenation," *IEEE transactions on neural networks and learning systems*, 2019.
15. J. Redden, "Predictive analytics and child welfare: Toward data justice," *Canadian Journal of Communication*, vol. 45, no. 1, 2020.
16. H. Ali, M. Khursheed, S. K. Fatima, S. M. Shuja, and S. Noor, "Object recognition for dental instruments using ssd-mobilenet," in *2019 International Conference on Information Science and Communication Technology (ICISCT)*. IEEE, 2019, pp. 1–6.

Figures

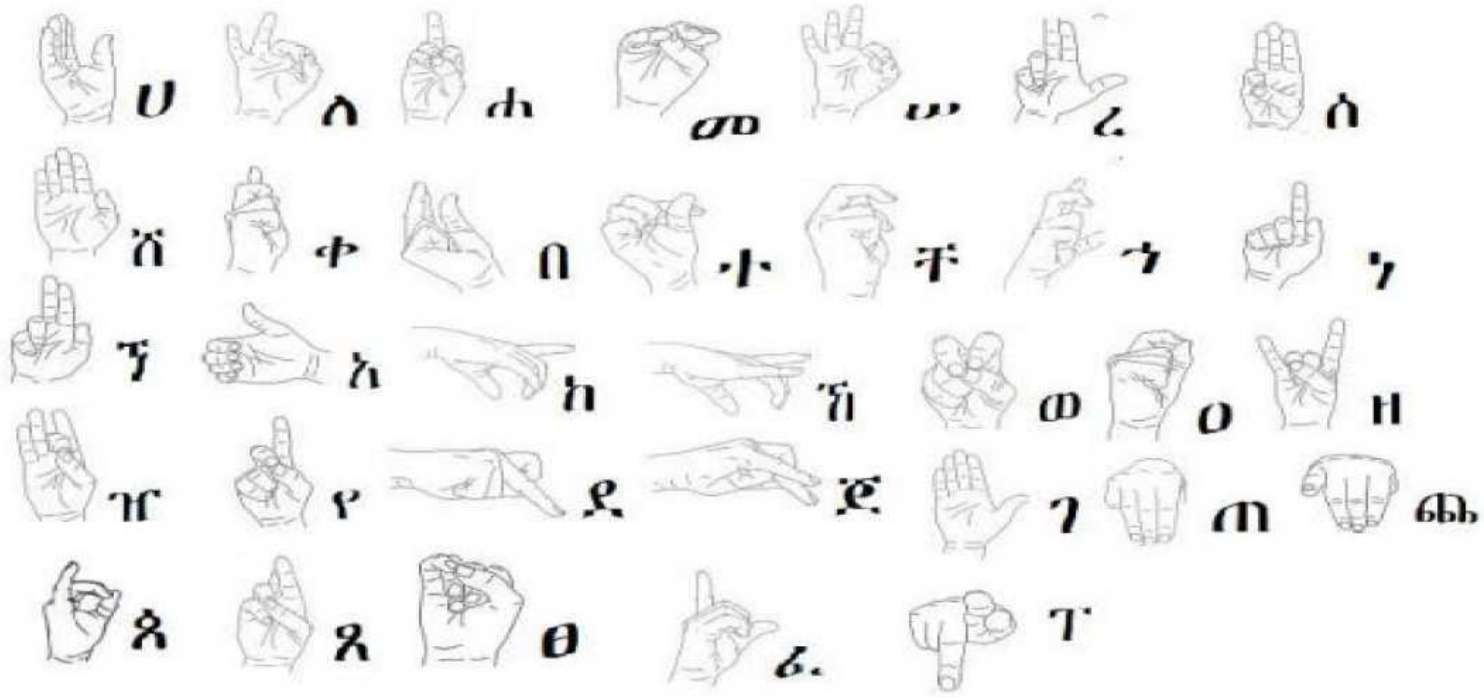

Figure 1

Amharic Sign Language Finger Spelling

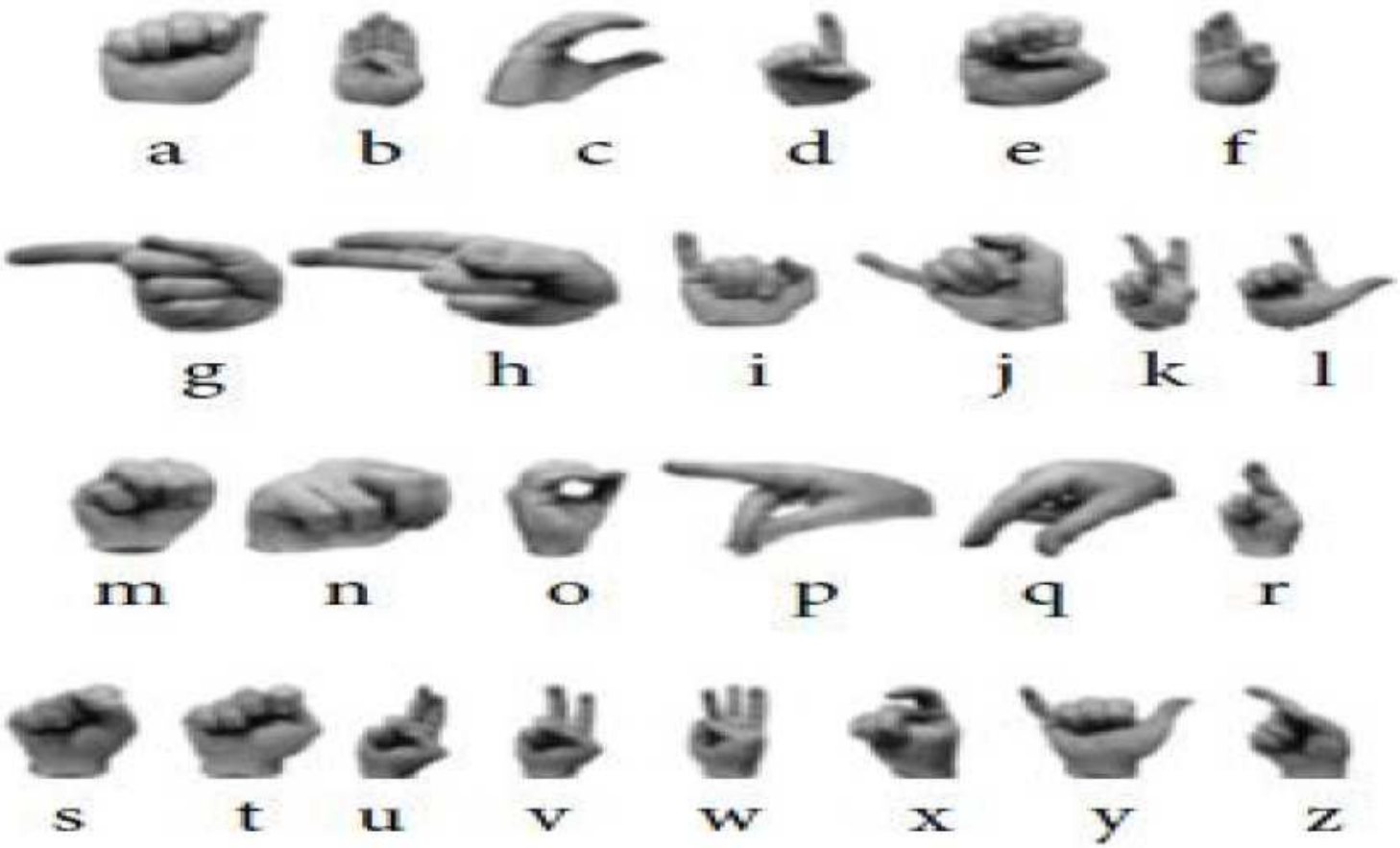

Figure 2

English Finger Spelling

Figure 3

Characters With Different Background

Figure 4

Character Recognition

Figure 5

Detection overlap example (it means 'Father') [3]

Figure 6

Extracted Frame Annotation Using LabelImg tool

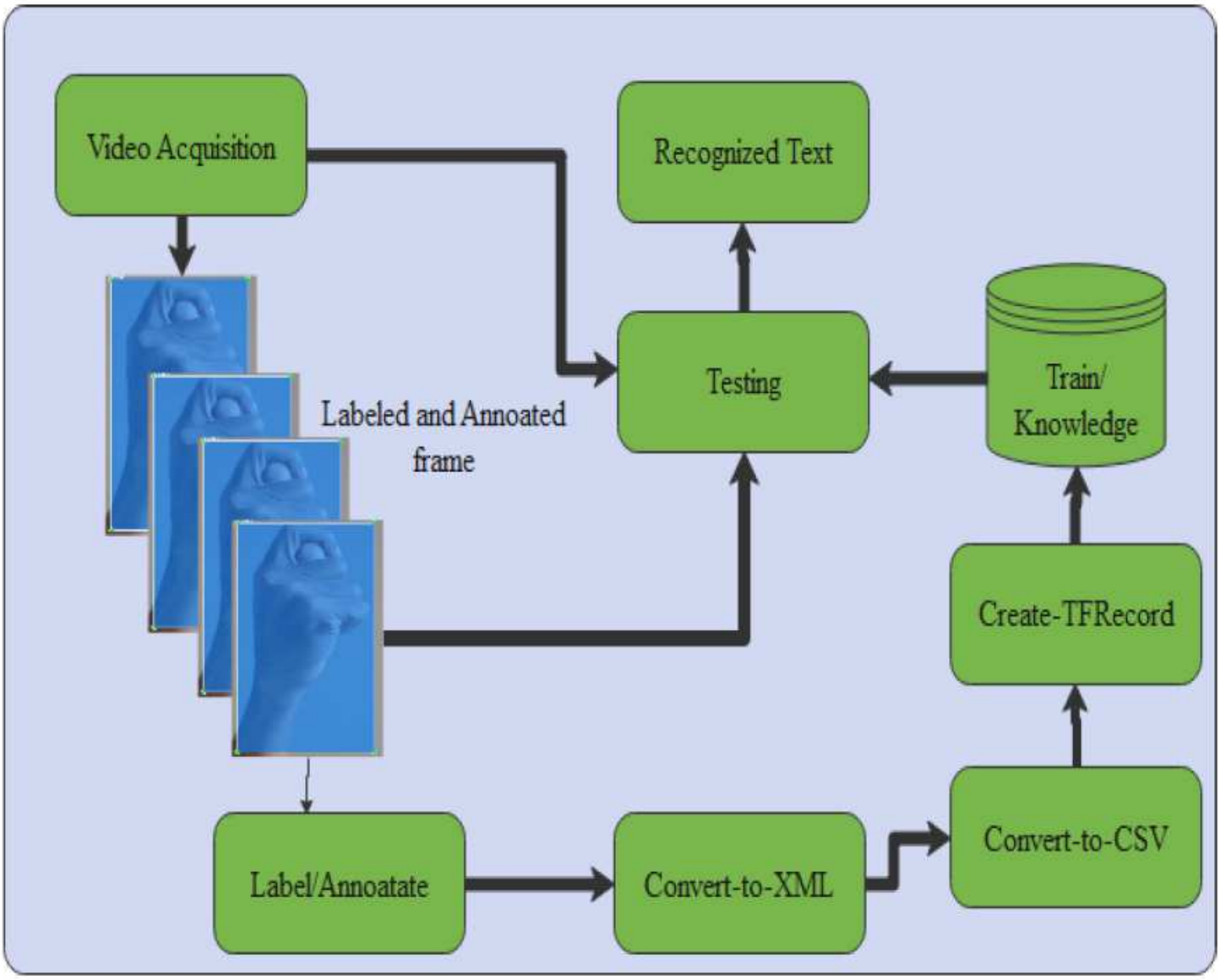

Figure 7

Data annotation and labeling

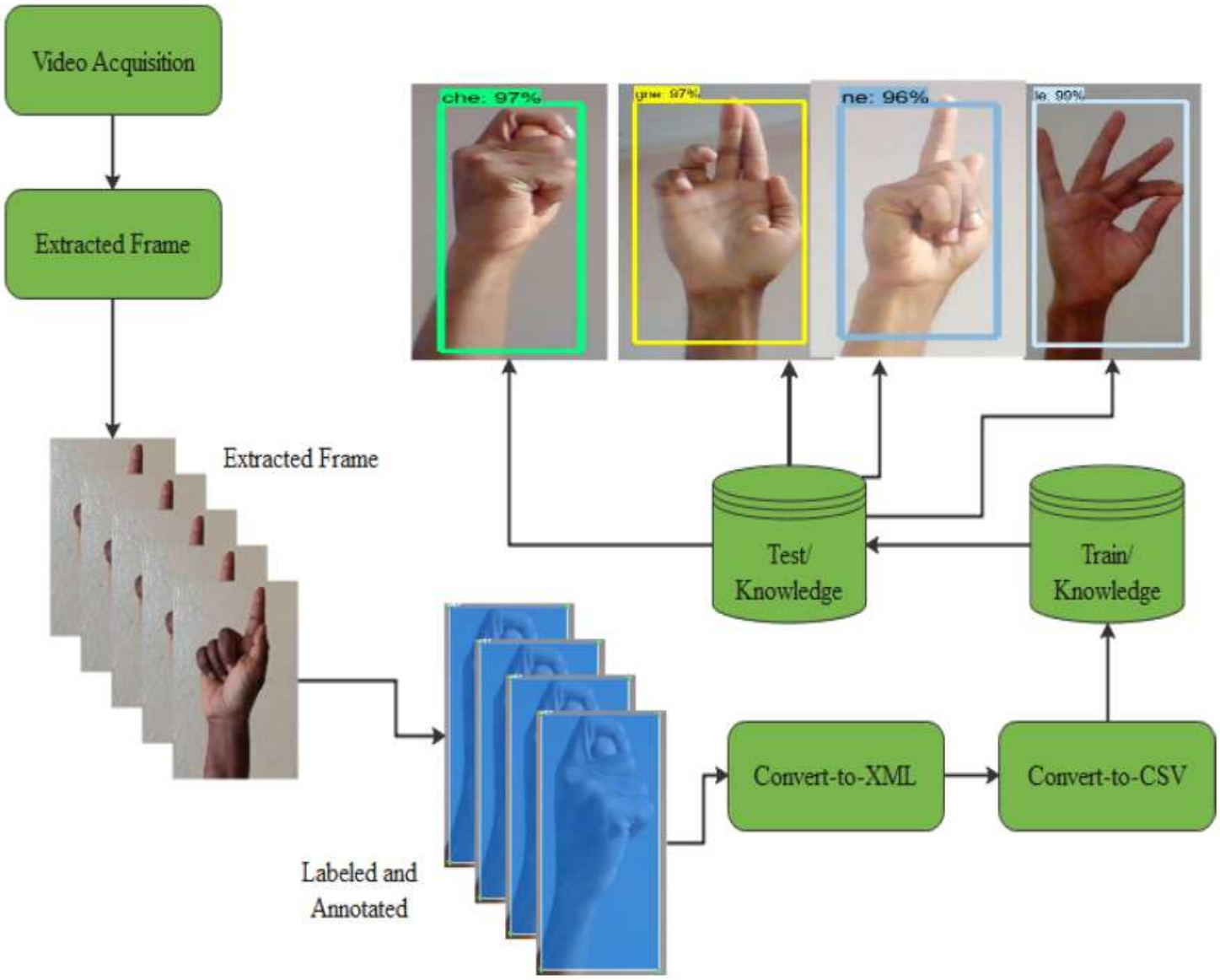

Figure 8

Recognized Character Using Faster-RCNN

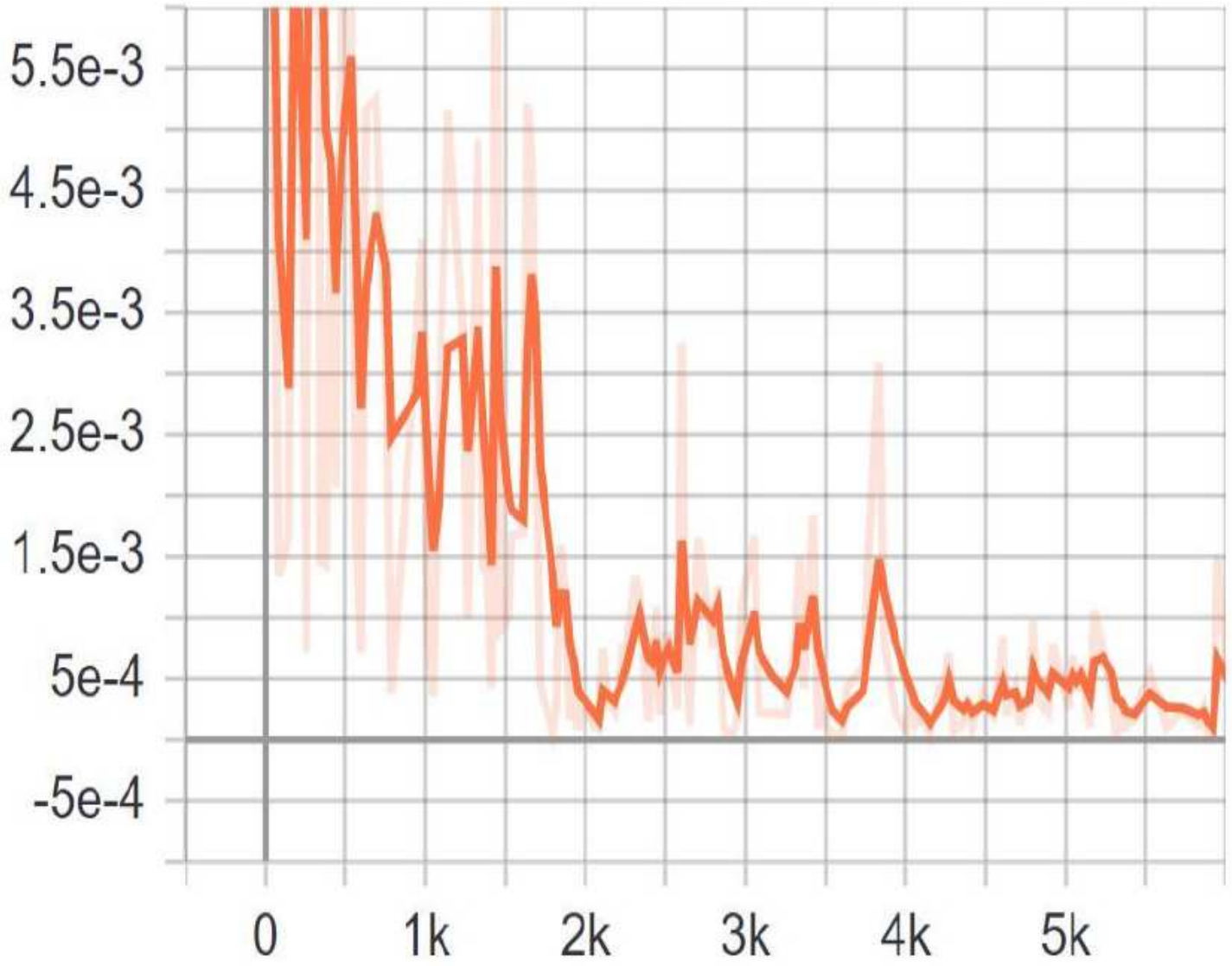

Figure 9

Loss Function of Faster-RCNN

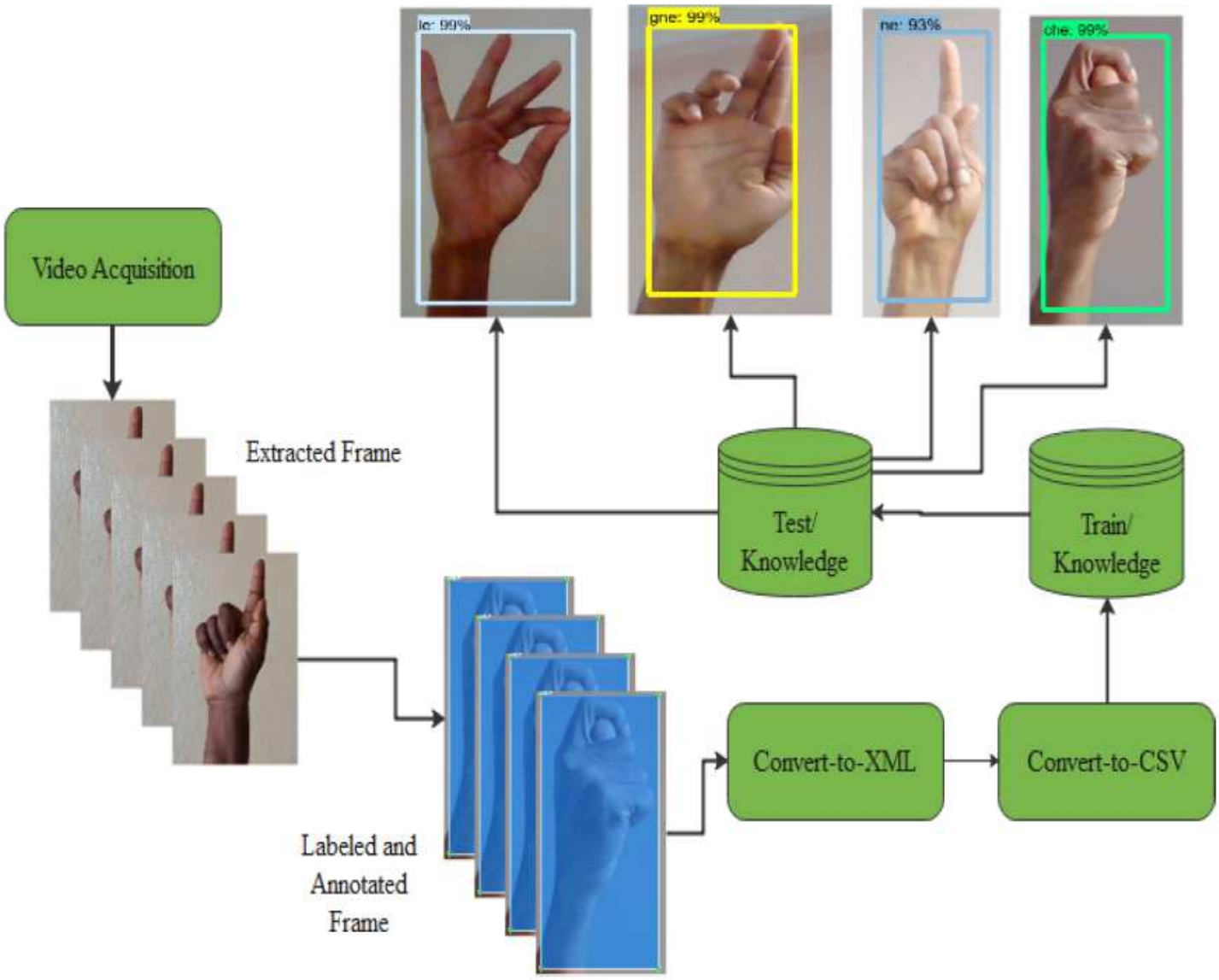

Figure 10

Recognized Character SSD