APS360 PROJECT PROPOSAL: ASL TO TEXT

Aashir Meeran
Student# 1007996082
aashir.meeran@mail.utoronto.ca

Shivang Mistry
Student# 1008151704
shivang.mistry@mail.utoronto.ca

Steven Li
Student# 1008025325
jilai.li@mail.utoronto.ca

Eugene Lee
Student# 1007932604
eug.lee@mail.utoronto.ca

ABSTRACT

With the rise of equity in North American society, as well as other parts of the world, American Sign Language has developed more importance than previous years. However, since the understanding of sign language often takes years to master, easier ways of interpreting such gestures need to be implemented in order to increase the accessibility of those who rely on ASL to communicate daily. This project aims to translate ASL into common English, using a combination of gesture recognition and deep learning training. Using PyTorch and OpenCV libraries, the final product should be able to recognize hand gestures and generate the meaning of the gestures in comprehensive, common English. The proposed timeline of this project is around 2 months.

—-Total Pages: 6

1 INTRODUCTION

American Sign Language (ASL) is the visual language used by the majority of individuals who are deaf or hard of hearing in North America. Unfortunately, most of society who is unaffected do not know how to interpret American Sign Language. In addition, existing tools for interpreting ASL are not widely available and affected individuals usually have to rely on human translators. We propose a project to address this issue by leveraging deep learning to develop an ASL to text converter computer vision software.

The aim of our project is creating a real time computer vision system to recognize ASL alphabet hand gestures and translate them to text. The system would involve taking real time camera feed and classifying corresponding letter which will be displayed in a simple interface. By enabling classification of real time gestures, it allows affected individuals to communicate more easily with non affected individuals.

This project is important and interesting as it has large potential to increase accessibility for deaf and hard of hearing individuals.Communication is an essential part of people's lives and helping bridge the communication gap between those who can interpret ASL and those who cannot will enable a more fair and inclusive society. In addition, the system can potentially be used in various public settings to enable communication between affected and unaffected individuals.

We have decided that this project warrants a deep learning approach as it has been proven to be very effective for computer vision tasks including object detection, segmentation and image classification. Deep learning models can be trained to classify complex images and therefore should be able to accurately recognize ASL gestures. We also made that there were datasets readily available as with Deep Learning projects, quality training data is essential.

To summarize, our project will be to develop a real time ASL to text converter system that recognizes gestures through a camera feed and translates them to text. We aim to tackle the issue of accessibility

in the deaf community and enable readily available and easy communication between hearing and non hearing individuals. This will be done by leveraging deep learning to develop a highly accurate and efficient system to be potentially applied in various public or industrial settings.

2 ILLUSTRATION

The illustration contains a rough idea of the model's architecture and a visualization of the interface, where the model detects the location and letter of the gesture along with a certainty rate.

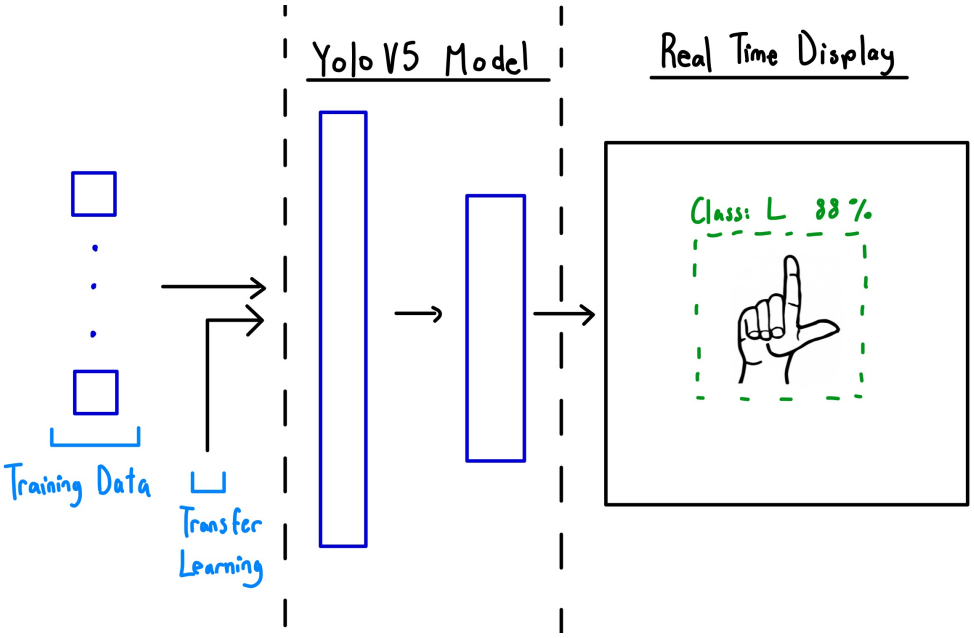

Figure 1: Illustration of the model

3 BACKGROUND AND RELATED WORK

Recently, in the field of deep learning and AI there has been a growing interest in using deep learning techniques to create computer vision systems for American Sign Language recognition. Here, i present one related paper and one existing software with a brief overview of how the two have contributed to the related space.

"Real-Time American Sign Language Recognition using Convolutional Neural Networks" by Fernando et al. (2018) is a paper where the authors proposed a CNN based approach for recognizing ASL gestures using a camera in real time. They ended up achieving an accuracy close to 95% with a data set consisting of 25 different gestures. This work evidently highlights the potential for applying a deep learning approach to the problem we are attempting to solve and provides meaningful context for our project.

Next is the "ASL ALphabet Recognition" software developed by a sign language converter team at Carnegie Mellon University. This software recognizes ASL alphabet and translate it to text using a similar CNN based approach as mentioned in the previous paper. In this case, a Hidden Markov model is also used to recognize gestures as it is effective at differentiating between transition periods between gestures and can isolate points in time where a gesture is intentionally made. Achieving an accuracy of close to 99%, the software provides a good benchmark for our system and demonstrates the potential for real world application.

4 Data Processing

Data processing and data preparation are crucial steps when developing ML projects, hence we plan on allocating a significant amount of time to this step. We will be using the American Sign Language Letters dataset shared by David Lee on the public dataset-sharing site, Roboflow. The dataset is minimal with only 720 images and annotations. There are 26 classes, a class for each of the 26 letters. The classes are relatively balanced with on average 28 images per class, as can be seen by the class distribution in Figure 2. Therefore, significant class balancing will not be required.

Figure 2: Class Balance of the Dataset

The cleanliness of the dataset will further be investigated by looking for and removing incorrect annotations, duplicate images, low-quality images, and corrupted files.

Object Detection tasks usually require a high volume of data, and clearly, our dataset is quite small. Hence, to increase our volume, we propose collecting our own data and adding it to the existing dataset. Furthermore, we hope to perform data augmentation techniques including rotation (+/- 10 degrees), shearing (+/- 10 degrees in horizontal and vertical), horizontal flipping, and introducing gaussian blur (2.5 px). For each data sample, about 4-5 augmented versions will be generated. Using these transformations, we hope to increase the training dataset to about 3500 images and reduce overfitting.

Next, the images will also be resized such that the images are all one size when input into the model. Furthermore, the images will be cropped to the object of interest. As discussed by Xu et al. (2018), detection can be improved by cropping and resizing images. Hence, we hope to achieve similar results with our project as well.

Lastly, we will perform normalization, by normalizing the pixel values of the images to a common scale of [0,1]. This is a common practice when training any sort of model, as it can help further stabilize the training process and improve accuracy.

The dataset will have an 80-10-10 split, where 80% will be training, 10% validation, and 10% test data. The majority of the dataset is used for training to reduce overfitting and improve model performance.

5 Architecture of Model

For our ASL to text converter project, we plan on applying transfer learning with the YOLOv5 machine learning network as our pre-trained mode. YOLOv5 is a well known object detection model developed by Ultralytics which has consistently achieved high accuracy on various computer vision

object detection tasks. OLOv5 stands for "You ONly Look ONce version 5", it is a convolutional neural network (CNN) architecture.

The architecture of the model is quite complex and beyond the scope of this course however I will summarize the key components including the backbone network, neck network and head network.

The backbone network is a CNN that is used to extract features for the image input. The neck is used to combine these extracted features and perform spatial pooling. SPatial pooling is done in CNNs to reduce the dimensions of features mapped by the network and prevent overfitting and increase the model's ability to generalize. Lastly, the head network consists of several more convolutional layers that is used to predict bounding boxes of objects and the classification probability of the objects in an image.

To adapt this model to our ASL to text converter, we will be applying transfer learning using a pre-existing dataset containing images of ASL hand gestures. We also plan on fine tuning the parameters of the pretrained model while re-training the model to recognize ASL hand gestures. This approach allows us to leverage the existing capabilities of state of the art object detection and focus on training the model to accurately perform our given task.

6 BASELINE MODEL

Nowadays, object detection is primarily done using deep learning. However, machine learning algorithms can still be used for such a seemingly complex task. The baseline model used is a combination of Histogram of Gradients and an SVM, which was developed by Dalal & Triggs (2005).

After performing data processing, features from the data need to be extracted. There are various feature extraction techniques, however, we will use one of the more common ones, the Histogram of Oriented Gradients. Once the features have been extracted, an SVM will be trained on the various HOGs for different classes. Then, the sliding window technique will be performed over the entire image, extracting features for each individual slide. The sliding window can have different sizes and aspect ratios to identify object locations. The features of each slide will be classified using the SVM. Finally, once the object has been classified, it will be localized using a bounding box, that can be easily determined since the coordinates of the sliding window will be recorded for each slide.

7 ETHICAL CONSIDERATIONS

Considerations must be made in regards to the diversity and acquisition of training data as well as consequences of errors or misclassifications in ensuring that ethical issues do not arise.

Namely, the model must be equitable, in that it is equally effective for all demographics. If the model fails to recognize gestures made by users of certain ethnicities, it could result in unintended harm. For this reason, it is important for us to ensure that our training data is diverse.

If we choose to collect data for the purpose of training this model, it is important to obtain informed consent from participants who provide data and to ensure their privacy by only using the data for the intended purpose.

If this model were to be released for public use, the potential harms caused by errors or misclassifications must be considered. The model should be tested thoroughly to ensure that it has a high accuracy rate. This means that in addition to promoting ethnic diversity in the datasets, training data should also include various lighting and background conditions, as well as variations in the gestures. The parameters should be optimized to prevent overfitting and allow for maximum accuracy on the test data.

8 PROJECT PLAN

Since the project is just initialized, not too much has been accomplished by any individuals. Instead, the team has worked together to come up with this project proposal. A rough work division plan can be seen below:

Table 1: Rough Work Division Plan

NAME	TASKS
Aashir	Model/training
Shivang	Model/training
Eugene	Real time Image from webcam using OpenCV and data processing
Steven	Data processing and hyperparameter tuning

A more detailed plan can be found on our Gantt chart. Moreover, a Google Calendar has been created so the team can track tasks more easily.

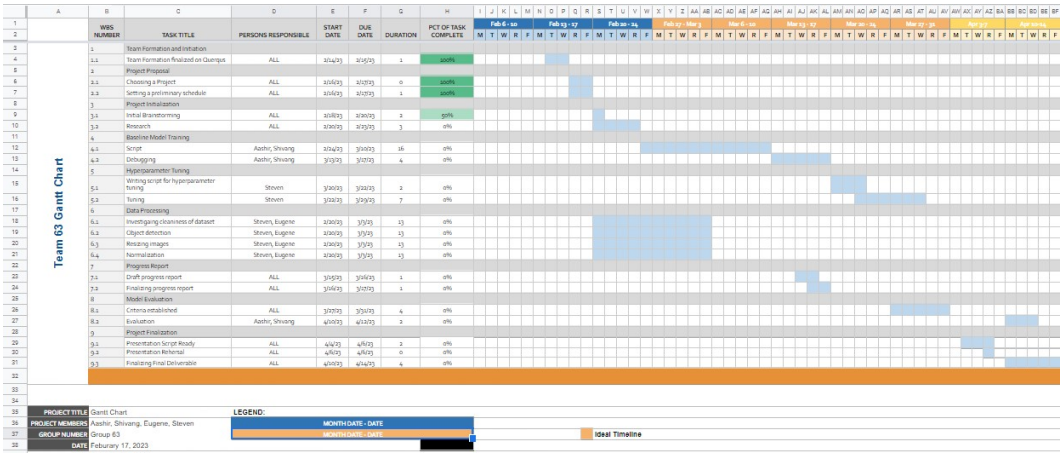

Figure 3: Snip Shot of the Gantt Chart

The team plans to meet bi-weekly on Discord, and bi-weekly in-person. The tentative time for either is on Thursdays 9 PM -10 PM and Fridays 11 AM -12 PM, respectively. During the online meetings, team members are encouraged to have the following points prepared:

- What they have accomplished in the previous week (w/ code);
- If they are stuck, how may the other team members help;
- Should a major decision come up, the proposing team member needs to ask all members for a vote;

If an in-person meeting is happening in a week, there is no need for the online meeting. However, team members also need to prepare the aforementioned articles. In addition to attending the meetings, team members are highly encouraged to be responsive and respectful on the Discord group chat. It is also suggested that when it comes to important matters, no one should ghost (be unresponsive on purpose) on the subject. If in emergency, a Discord voice call may be implemented with sufficient rationale. The team will use a combination of Google Collaborations Notebook and GitHub to make sure the code is not overwritten. In addition, since the members are required to attend weekly team meetings with a brief description of what they have accomplished, the chance of overwritten code will be adequately reduced.

9 Risk Register

Project completion risks are events that are to be taken seriously. It is important to identify how to mitigate these risks in order to ensure the project is completed smoothly without significant setbacks. The following are some of the major risks and how we will address them:

1. Insufficient Training Data: A major risk of a deep learning project especially one involving classification Is not having enough training data to allow the model to generalize towards cases it has not seen. Another potential issue would be not having enough quality data points to train the model on. Lastly the data could be biases which may lead to inaccurate results or the model being unable to make accurate classifications on new data. This risk has been determined to be moderate as we have looked into various data sets and found a suitable ASL dataset for transfer learning which contains high quality standardized images. We will also be generating some of our own training data to perform preprocessing and cleaning which will ensure the data is unbiased and consistent.

2. Model Overfitting: A second risk for this project is the potential for the model to overfit to the training data. Overfitting is the result of a model becoming too specialized to recognize the training data and thereby being unable to generalize to new testing data. This is a high level risk as we will be using a pretrained model and the transfer learning will occur with a relatively small dataset. To mitigate this risk we will generating our own dataset in addition to the online one we found to increase the variability of training and test data. We will also apply early stoppage technique to stop the model early when the testing accuracy stops increasing.

3. Hardware Limitations: For any deep learning project, hardware is always a consideration as it is the limiting aspect of the training aspect. For this project this risk is deemed to be low as we will primarily be working off a cloud server with Google Collab. however cloud server speeds may also be limiting as you are only allocated a set amount of memory capacity. If this issue were to arise, we could consider reducing the complexity of the model, reducing the size of training batches or reducing the number of training epochs.

4. Not being able to complete deliverable on time: Given such a short period of less than 2 months, there is a certain risk of not being able to complete this project on-time. To counter this risk, we will use a combination of Google Calendar, Gantt Chart and Trello to track the due dates. Team members are to check the Trello page every week to make sure no deadline has been missed. Trello is chosen because it is a website built for project management.

10 Link to Github or Colab Notebook

`https://github.com/lijilai1313113/APS360-Project`

11 Citations, figures, tables, references

References

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

Jianfeng Xu, Lertniphonphan Kanokphan, and Kazuyuki Tasaka. Fast and accurate object detection using image cropping/resizing in multi-view 4k sports videos. In *Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports*, MMSports'18, pp. 97–103, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359818. doi: 10.1145/3265845.3265852. URL `https://doi.org/10.1145/3265845.3265852`.