Bachelor Project - Sign Language

Tobias, Bertram, Silas, Patrick

March 25, 2024

Abstract

1 Abstract

Your abstract.

2 Introduction

ASL is a abbreviation for American Sign Language. Sign means a static placement of fingers and hands to illustrate a word or a letter. Gesture means a sign that includes movement.

2.1 Sign Language & Machine Learning

2.2 ASL Alphabet

The ASL alphabet is a way to communicate letters through Sign language. It consist of 24 static signs and two signs that require motion.

3 Dataset

3.1 Letters A-I & K-Y

We use a collection of static ASL images[Nag18] to train the letters A-I and K-Y as these signs do not require movement and can be processed as a single image.

The training data set contains 87,000 images which are 200x200 pixels. There are 29 classes, of which 26 are for the letters A-Z and 3 classes for SPACE, DELETE and NOTHING.

This dataset only consist of static images for the letters J and Z and since these letters consists of movement the dataset is not comprehensive enough.

3.2 J & Z

We use a collection of ASL videos for the letters J and Z as their meaning can only be conveyed through movement. These videos are sourced from Kaggle Dataset [TEA21]. This collection contains 316 and 396 video's of a variety of people performing the sign for the letters J and Z respectively.

4 Method

4.1 Media Pipe

A landmark is a MediaPipe term for a feature of a hand. In the world of MediaPipe there are 21 landmarks, one for each joint as seen on fig. 1.

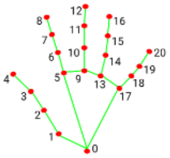


Figure 1: Enter Caption

Landmark extraction happens through the use of the MediaPipe Hands library[ZBV+20]. The extracted landmarks are then processed in one of two ways depending on the model to be training, either a *dynamic* or a *static* model.

4.2 Static gesture data preparation

The landmark data extracted from the static gesture data corresponds to screen coordinates. However, our primary concern lies not in the hand's placement on the screen, but rather in the relationship between landmarks relative to one another on the same hand. Thus, prior to preparing the data for our model, we normalize the landmarks to fall within a range of -1 to 1. In this normalization process, landmark 0 is assigned the value of 0, see fig. 1.

4.3 Dynamic signs

In order to train a model on signs which contain movement, some further processing had to take place. We will here explain the steps from fig. 4.

4.3.1 Trajectories

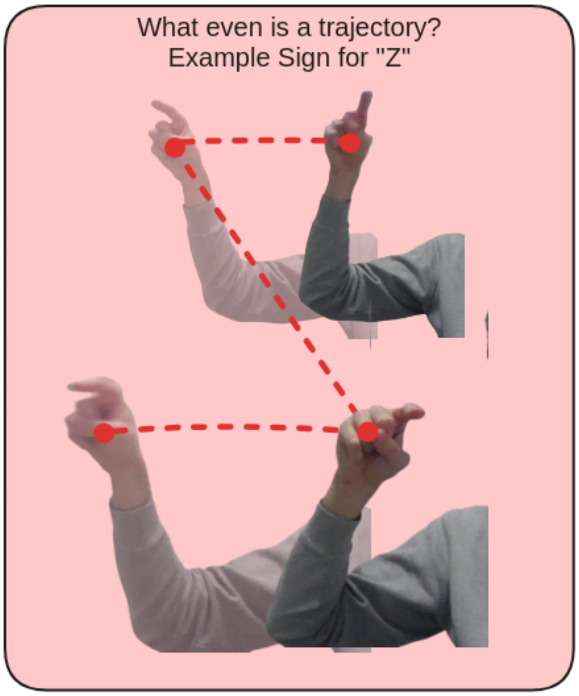


Figure 2: An illustration showing the idea behind calculating the trajectories of a video.

A trajectory element consists of three values corresponding to a direction in 3D space based on the difference in x, y and z coordinate means from one frame to the next. Possible values for a direction:

- 0: Stationary
- 1: Up, Right, Into
- -1: Down, Left, Away

4.3.2 Extracting Keyframes

A significant problem arose when dealing with video data as opposed to static image data: Videos are most often of varied lengths. As of now, our model does not support receiving a varied number of features. As such, we need to extract some target number of frames from all videos. This number needs to satisfy the model's expectation of a fixed number of features. The function that extracts this number of frames needs to ensure quality of the data.

We solved this by always picking the first and last frames, and randomly selecting elements from the rest.

5 Features

5.1 Landmarks & Trajectories

We wish to accomplish 2 things by supplying the model with both landmarks and trajectories:

- Recognition and precision in dynamic signs: Inclusion of trajectories along with the landmarks allows the model to distinguish more easily between signs which may not vary much in terms of the position of the hand, but varies in movement.
- Using a single model to recognize both static and dynamic signs: The model will be trained on data which, for static signs, contains trajectory data with 0s (stationary) exclusively. For this reason, the model will have to rely almost entirely on the landmarks for recognition of these signs, which eliminates the need to train a separate model for static landmarks only.

5.2 Static model

Trained with 21 features, one for each landmark, using trained using Softmax Logistic regression.

5.3 Dynamic model

For the Dynamic model we currently use 3 frames each with 21 landmarks as features. We also get features from the trajectory that is x,y,z coordinates from all the frames except the first one. Hähem, meaning that the model expects a total of $(k-1)*3+f*21*2$ features or 132.

6 Results

See results for static gesture recognition: confusion matrix fig. 5 and classification report 1. Results for dynamic model is work in progress.

7 Introduction

8 Dataset

Static

9 Static gesture features

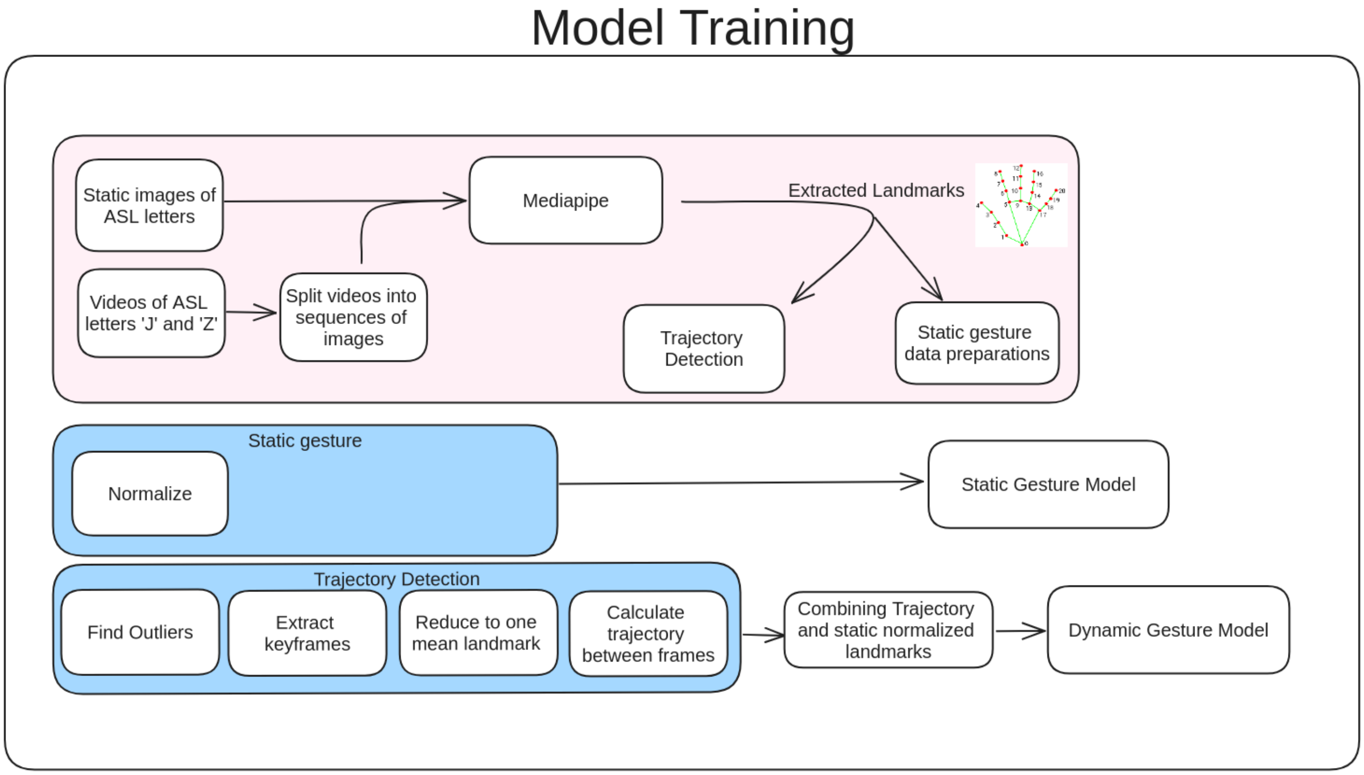
- 21 * 3 features

10 Dynamic gesture features

- 12 * 22 + ?

11 Appendix

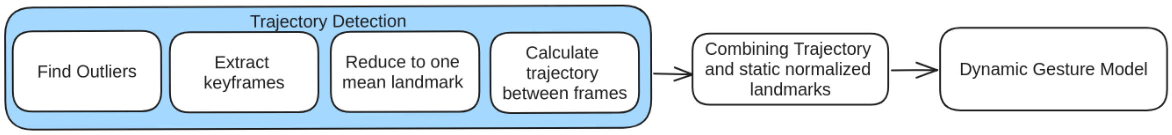
11.1 Overview of the model-training-step



Model Training
Static images of ASL letters
Videos of ASL letters 'J' and 'Z'
Split videos into sequences of images
Mediapipe
Extracted Landmarks
Trajectory Detection
Static gesture data preparations
Static gesture
Normalize
Static Gesture Model
Trajectory Detection
Find Outliers
Extract keyframes
Reduce to one mean landmark
Calculate trajectory between frames
Combining Trajectory and static normalized landmarks
Dynamic Gesture Model

Figure 3: An overview of how we train our Sign Language recognition model. The pink box shows how raw image and video data is processed into landmarks. The blue boxes illustrate how landmarks are pre-processed and fed to a model for training.

11.2 Preprocssing of data for dynamic gesture



Trajectory Detection
Find Outliers
Extract keyframes
Reduce to one mean landmark
Calculate trajectory between frames
Combining Trajectory and static normalized landmarks
Dynamic Gesture Model

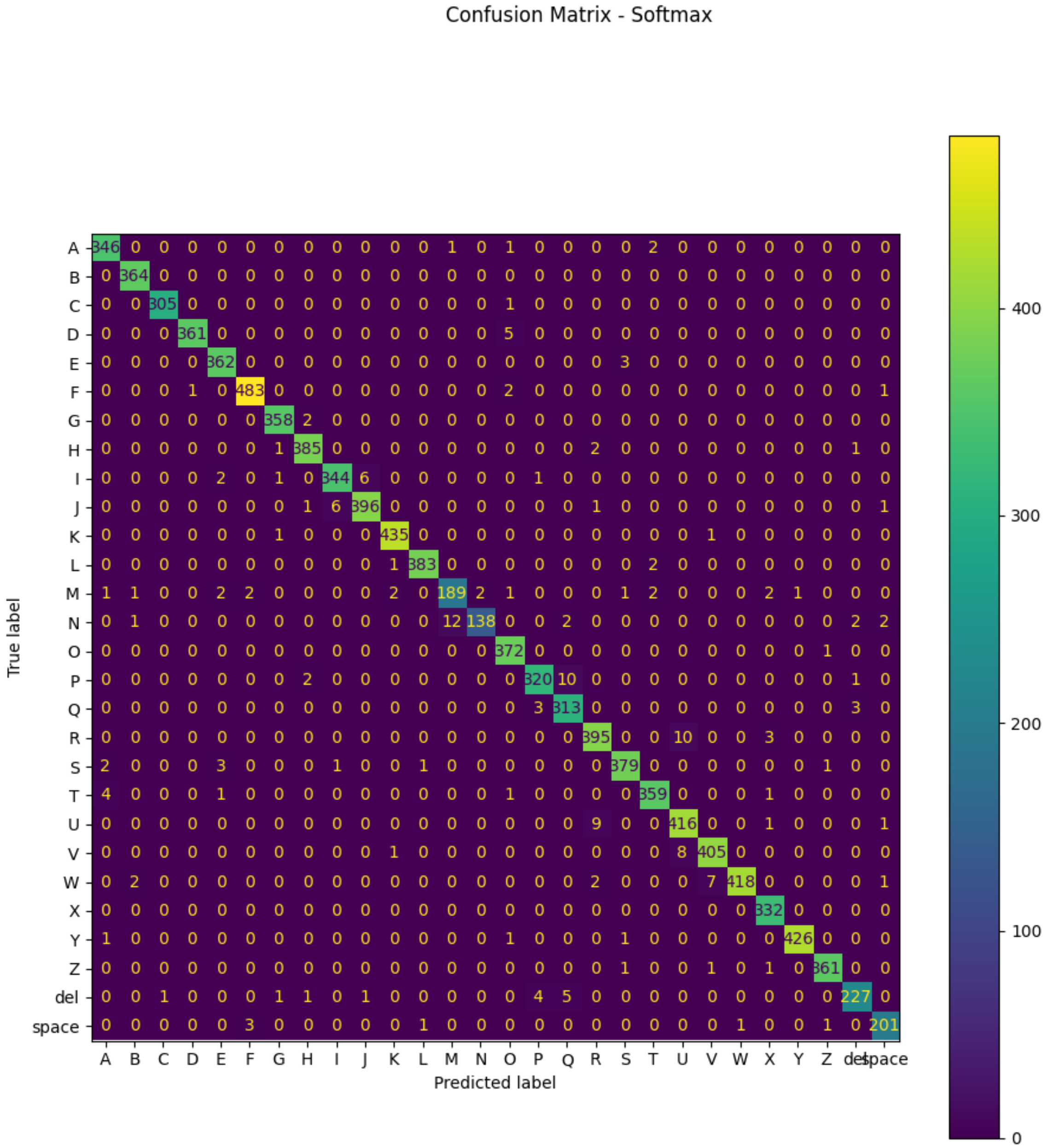
Figure 4: The steps for preprocssing data to train a model for dynamic gestures

11.3 Static gesture model classification report

Class	Precision	Recall	F1-score	Support
A	0.94	0.97	0.95	31
B	1.00	1.00	1.00	39
C	0.98	0.98	0.98	43
D	1.00	1.00	1.00	38
E	0.95	1.00	0.98	40
F	0.98	0.98	0.98	49
G	1.00	1.00	1.00	40
H	1.00	1.00	1.00	40
I	0.98	0.93	0.95	45
J	0.97	0.97	0.97	34
K	1.00	0.97	0.99	35
L	1.00	1.00	1.00	38
M	0.92	0.69	0.79	16
N	0.93	0.82	0.88	17
O	1.00	1.00	1.00	42
P	0.97	0.95	0.96	41
Q	0.86	1.00	0.92	24
R	0.90	0.97	0.93	36
S	0.98	1.00	0.99	44
T	0.97	1.00	0.99	35
U	0.95	0.90	0.93	42
V	0.96	0.98	0.97	45
W	1.00	0.97	0.98	33
X	0.90	1.00	0.95	27
Y	1.00	0.98	0.99	44
Z	0.97	0.94	0.95	33
del	0.96	0.89	0.93	28
space	0.95	1.00	0.97	18
Accuracy			0.97	997
Macro avg	0.96	0.96	0.96	997
Weighted avg	0.97	0.97	0.97	997

Table 1: Classification Report for our static model

11.4 Confusion matrix for static gesture model



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	del	space
A	346	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0
B	0	364	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	305	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	361	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	362	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
F	0	0	0	1	0	483	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1
G	0	0	0	0	0	0	358	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	1	385	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0
I	0	0	0	0	2	0	1	0	344	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	1	6	396	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
K	0	0	0	0	0	0	1	0	0	0	435	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	1	383	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
M	1	1	0	0	2	2	0	0	0	0	2	0	189	2	1	0	0	0	1	2	0	0	0	2	1	0	0	0
N	0	1	0	0	0	0	0	0	0	0	0	0	12	138	0	0	2	0	0	0	0	0	0	0	0	0	2	2
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	372	0	0	0	0	0	0	0	0	0	0	1	0	0
P	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	320	10	0	0	0	0	0	0	0	0	0	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	313	0	0	0	0	0	0	0	0	0	3	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	395	0	0	10	0	0	3	0	0	0	0
S	2	0	0	0	3	0	0	0	1	0	0	1	0	0	0	0	0	0	379	0	0	0	0	0	0	1	0	0
T	4	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	359	0	0	0	1	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	416	0	0	1	0	0	0	1
V	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	8	405	0	0	0	0	0	0
W	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	7	418	0	0	0	0	1
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	332	0	0	0	0
Y	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	426	0	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	361	0	0
del	0	0	1	0	0	0	1	1	0	1	0	0	0	0	0	4	5	0	0	0	0	0	0	0	0	0	227	0
space	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	201

Figure 5: Confusion Matrix Softmax multiclass classification

References

[Nag18] Akash Nagaraj. Asl alphabet. https://www.kaggle.com/dsv/29550, 2018.

[TEA21] SIGNN TEAM. Asl sign language alphabet videos [j, z]. https://www.kaggle.com/datasets/signnteam/asl-sign-language-alphabet-videos-j-z, 2021.

[ZBV+20] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. MediaPipe Hands: On-device Real-time Hand Tracking, June 2020. arXiv:2006.10214 [cs].