Utilizing Computer Vision to Detect Grammatical Facial Expressions of American Sign Language

Sasha Malysheva mentored by Karen Ge

August 1, 2020

Abstract

In this paper we focus on the grammatical facial expressions of ASL, expressions that denote the meaning of a sentence. Through the use of computer vision libraries we were able to track the coordinates of various facial landmarks that denote expression. This information was then used for a logistic regression to classify images into their grammatical facial expression category. This project focused on negative sentences, wh-question sentences, and yes/no question sentences. We were able to get an F1 score of 0.96103 for our test dataset, as well as a precision of 0.96065 and recall of 0.96333. This work shows that facial landmark tracking technology can be utilized in improving machine based ASL translation. American Sign Language translation is incredibly important for both increasing accessibility for the hearing impaired and deaf community as well as for promoting widespread ASL education.

1 Introduction

Approximately 37.5 million Americans over the age of 18 report trouble hearing, accounting for 15% of the adult population, yet American Sign Language (ASL), the primary signed language in the United States (and portions of Canada) does not have a standardized dictionary. [24]Although estimations for the total population that speaks ASL as their primary language are outdated and inaccurate due to the lack of ASL representation on the census, speculations have been made in 2006 that it is about 500,000.[18][24] American Sign Language is it's own language with a linguistic, phonetic, and grammatical structure completely different from English. The visual nature of the language introduces a multitude of difficulties when attempting to translate it mechanically, as typical natural language processing techniques are nullified in the face of ASL. In order to truly translate ASL, one must take into account all the intricacies of the language. Current technologies that attempt to translate ASL heavily prioritize the signers hands, effectively disregarding a large portion of the language that comes through non-manual markers.

[].

2 Literature Review

In order to qualify as a true translation the five key parameters of ASL articulation must be taken into account [13]. The parameters are: handshape, movement, place of articulation, hand orientation, and non-manual markers. The non-manual markers, which are also known as suprasegmental features, are the most overlooked in the current 'translation' models, but are often crucial to the understanding of ASL [13]. These features include head movements, facial expressions, and the exaggeration of signs. Suprasegmental features are an integral part of the grammar of ASL. They are used to signify a question or the type of question; they are used as a replacements for certain words, for example the word 'very' in sign language is expressed as a different facial expression. They are used to convey the connotation of the word, the way tone is used in spoken language. Furthermore, the level of exaggeration or size of the sign can signify the same thing that volume/emotional tone signifies in spoken language.[16][21][8][1][2][3]

When speaking any language, one will express themselves through their facial expressions; in signed languages such as ASL, the facial expression takes on a greater role that is beyond simply conveying emotion and is grammatically part of the language. In these grammatical facial expressions, the top part of the face - eyes, eyebrows, and in some cases the nose- conveys sentence type and other grammatical information. The lower half of the face - mouth and lips - is utilized to communicate descriptive information through specialized sets of movements. The information of the mouth and lips describes numerous details about what is happening in the sentence, such as the size of an object, emotions, 'volume', etc.[16][21][8][1][2][3]

In (source) it lists the following as sentence types: declarative, imperative/command, negative (i.e. negation), topic-comment, question (wh- question, yes/no question), rhetorical questions, and conditional. Each of these sentences has a certain set of non-manual markers which must accompany it being said.

[18]

2.1 Negative Sentences

"Negative sentences in ASL have the following features: shaking the head from side to side throughout sentence; possible frowning or squinting (eyebrows squeezed together); optional use of negation term: NOT, NEVER, NONE, CAN'T, DON'T, WON'T, etc.; often at the end of the sentence, or verb that incorporates the negative: DON'T-WANT, DON'T-LIKE, DON'T-KNOW."

2.2 Yes/No Questions

"... the following nonmanual signals must accompany a wh-question in ASL: eyebrows are furrowed or eyes are squinted; head tilts downward; body may lean slightly forward, with the shoulders raised; direct eye gaze at addressee last sign held - waiting for a response."

2.3 Wh- Questions

"In yes/no questions in ASL, the following occur along with the signs: raised eyebrows, eyes widened; head or body tilted forward; sometimes the pronoun or the sign glossed as HUH or '???' is added at the end of the sentence; the last sign is held longer sometimes; the movement of the last sign is repeated."

Classifying the type of sentence is especially important in signed languages such as ASL because the same string of words can mean different things based on the grammatical facial expression given. For example, the signed words that translate to, "I am good" change their meaning to "I am not good" with the negation facial expression. Similarly, the same string of signed words, "Sasha know Karen" that are signed in the same order can be expressed as a declarative statement, "Sasha knows Karen," question, "Does Sasha know Karen," or negative statement, "Sasha does not know Karen," simply depending on the expression.

A fundamental linguistic understanding of a language is crucial when pursuing a more holistic machine translation then is currently available. In addition, must consider the target demographic when working on a project that aims to help people, and one must ascertain that this technology will do what it is intended to do, without inhibiting the people that it aims to aid. Members of the deaf community have expressed criticism concerning some technologies that were created in an effort to increase accessibility by translating ASL [17][11][22]. Technologies such as 'translating gloves' have been critiqued for being too bulky, not reliable, and most importantly they did not take into consideration all the parameters of articulation discussed above - only sensing the hands; overall the gloves proved not to cater to the deaf community in the way that they were

intended [17]. The other demographic that has a use for an ASL translation model or even a standardized ASL dictionary is the population of people who are learning sign language. ASL translation models need to be developed to create an accessible place for sign language education and translation, especially because upward of 90% of deaf children are born to hearing parents and may not have access to sign language education from an early age. [14] Another layer to consider when exploring ASL translation and an ASL dictionary is the differences in ASL evolution between different races in the United States. ASL is a relatively new language, and it developed at a time when the U.S. was heavily segregated, and therefore there are considerable differences in the way black and white people were taught sign language. [12][29][7] Through a standardized dictionary that includes different racial and geological dialects, American Sign Language could grow more unified.

Machine models that currently translate ASL are in their early stages. The majority can translate the ASL alphabet into the English alphabet along with numbers, although more advanced models range from detecting a certain amount of gestures to being able to translate longer conversations with varying accuracy. These more advanced models often require many angles of cameras in order to detect the 5 parameters of speaking.[20][23] Currently little research has been done in grammatical facial expression recognition and classification specifically for American Sign Language. There have been several studies done on Libras (Brazilian Sign Language) such as a study conducted by University of California Irvine. [4]Another study that pertains to American Sign Language and grammatical facial expressions explores the animation of sign language. Animation of grammatical facial expression is particularly interesting because of its potential to increase online accessibility for people who are fluent in ASL but illiterate or unfamiliar with written English.[15] [19][28]

3 Purpose

1. Utilize facial landmark tracking techniques to accurately predict the grammatical facial expression of a person signing.

2. Aid the effort to create a machine learning model that has the ability to translate ASL, for the purposes of an online standardized dictionary, educational purposes, or translation purposes

4 Methods

This project chose to focus on only three types of sentences. This means that the project had three classes with one grammatical facial expression per class. The chosen expressions denoted the following types of sentences: negation sentences; who/what/when/where/which/how questions (w/h questions); and yes/no questions. These three categories were chosen because they are three very common sentence types for which grammatical facial expressions are very important (potentially more so than in sentence types like topic sentences), and because these facial expressions are among the most distinctive and disparate in the face. This means that these expressions can be identified more easily through their facial landmarks rather than their head and body positioning -

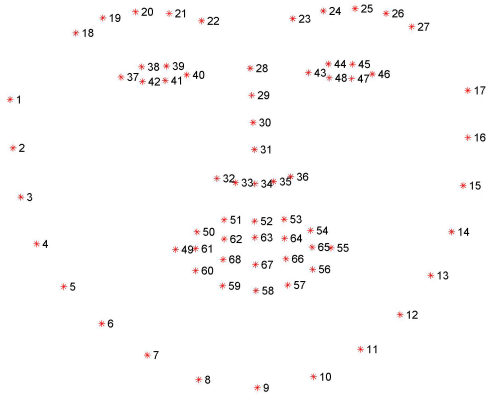

Figure 1: Depiction of the 68 coordinates in the *dlib* facial landmark dataset.[6]

unlike other grammatical facial expressions that depend more heavily on the movement of the body or contents of the sentence.

The model that we designed for this purpose was first fed a dataset of photos for each of the classes that we were examining. First, it completed some basic preparations for each photo in the dataset: resized it to 500, turned it to a grayscale image (from RGB) using OpenCV. Then it detected the region of interest and focused on the rectangular region of the image that contained the face. We used OpenCV and imutils - specifically the face_imutils submodule - in order to detect faces.[25][26][27][9] Next, the model was tasked with finding the coordinates of specific facial landmarks. This program operated with the help of OpenCV and dlib libraries and was adapted from (insert pyimage tutorial here). *dlib*'s facial landmark detection was pretrained on the $68 point iBUG300-W dataset$ inside the dlib library to locate $68(x, y)$ coordinates which map the facial structures of the face.[10][Figure 1] The program we designed collects these coordinates and adds them to a comma separated value file (CSV) to create new dataset. This function also adds an additional column which labels the data numerically for the class. After this function is executed on each grammatical facial expression class with a folder of images, the model moves on to the next function.

Next, each CSV file that is outputted from the previous is split into a train, validate, and test dataset at a $.8, .1, .1$ split respectively. Then the code moves onto a third function. In this step we combine the train, validate, and test subsets from each class into one train, validate, and test dataset each. This function is where the logistic regression will happen. From the library sklearn and its submodule *linear_model* we used the $LogisticRegression()$ class. Of the parameters that we passed through to this function, the most important was $multiclass = `ovr'$. This parameter allowed the logistic regression to perform it's 'One vs. Rest' function which fits a binary model for each label. In this case - with three classes - this creates three separate binary classification problems: negation vs. [w/h questions and yes/no questions], w/h questions vs. [yes/no questions and negation], and yes/no questions vs. [negation and w/h questions]. Another parameter that was passed in for the $LogisticRegression()$ class was the max_iter which is the "Maximum number of iterations taken for the solvers

to converge". For this parameter we found that the F1 scores with the default were approximately 0.9, and when altered the max_iter to 500 the $F1$ scores rose to .97 for train and 0.96 for validate. The final parameter that we experimented with is the penalty, which we kept at the default of '$penalty = l2'$.

Then each of these datasets are split into the coordinates, and label column. This step is completed because the $sklearn.linear\_model.LogisticRegression.fit(x, y)$ function takes in two parameters and they were the coordinates and label column respectively. The reason that the labels are numerically tagged and in one column is because: when they were categorically tagged with the name of their class, we needed to perform one hot encoding in order to get them back into numerical value. This process created an array that was three columns wide and the length of the training data set, which was not accepted into the y value of the fit function which only accepted a one dimensional array.

In order to retrieve an F1 score to gauge the accuracy of our model, we needed to utilize the $sklearn.linear\_model.LogisticRegression.predict(x)$ function, for which we passed in our training dataset of coordinates - the same data we used for the x of the fit function. The output of that step is our training dataset's $\hat{y}$ (i.e.yhat). Finally, an F1 score is found with the following function $sklearn.metrics.f1\_score()$ for which we passed in y_train (the y value from the fit function), $yhat\_train$ (established in previous line), and $average =$ '$macro'$ (this parameter is required for a multiclass model). '$macro$' was chosen because this parameter worked well for our model because it treated each '$label$' or class as equal and it 'Calculate metrics for each label, and find their unweighted mean' according to (apa citation for sklearn documentation page).

4.1 Data Collection

The desired data for this project was image frames taken while one person is signing something in ASL - our model is catered to only one person being in the photograph. From these frames our model would be able to detect the face of the signer, and perform the remainder of the experiment. In order to tag our data for our model, we need to know the type of grammatical facial expression demonstrated in the picture.

Due to the relative specificity of the data we were looking for, there were no datasets available in American Sign Language. Similar projects exploring grammatical facial expressions in signed languages have utilized a Libras sign language dataset, but considering our project was focused on the ASL implementation, this dataset did not fit our criteria. An alternative option was to clean video data from conversations in ASL that had been filmed, and extract specific frames. But in the timeframe of this project, the amount of data cleaning and manual tagging would have taken far too much valuable time currency. Finally it became clear that the data needed to be created.

The type of data we created aimed to prove if our model could classify and predict facial expressions in a very controlled environment. Therefore, we were adamant about keeping various factors the same to allow us to find the maximum accuracy of our model. To attempt to preserve the integrity of the data as much as possible there were several protocols and procedures established. The data was taken in the form of picture bursts from the web application Google Teachable Machine.[5] This program was chosen for a multitude of reasons: ability to take bursts of photos, ability to choose the number of frames per

second, automatically creating all the photos at a small size and relatively low pixel amount, and ability to classify bursts into folders and download folders as zip files.

The photograph bursts were taken at 24 frames per second in order to imitate the action of recording a video of a person signing and extracting individual frames. The images were square photos that are each 224 by 224 pixels. Each photo was taken in color. The photographs were taken in front of a white background - painting canvas. The subject (Sasha Malysheva - researcher) was wearing a black shirt and her hair always began in the same way behind her back, the hair did move around somewhat when the subject was signing but it didn't obscure the face with the exception of the ears. The computer from which the photos were taken was placed in the same spot for them all and did not move. The images were taken at the same time of day, and there was a light pointed at the subject and background from behind the computer, it was standing in the same position to make certain that the lighting and shadows were kept consistent from picture to picture.

The subject took approximately 1000 photos in each of the three classes. Specifically there were 1,007 photos in the negation class; 1,044 photos in the w/h questions class; and 1,001 photos in the yes/no questions class. The slight discrepancy in the amount of photos per class is a result of a slightly longer burst in some cases in order to finish a specific sign. During the data collection the subject was modeling the facial expression of the class, and its respective variations throughout the entirety of the burst. Variations on the expression are in several forms. Varying levels of severity - e.g. in the category of negation signing not good could be mildly not good or extremely not good. Differing emotions depicted through the grammatical expression - e.g. a sad 'not good' versus an angry 'not good'. The expressiveness and 'volume' -e.g. exaggerated 'not good'. Mouthing words or not mouthing them - e.g. saying 'not good' with a closed mouth or with an open one in order to mouth the words - this variation goes hand in hand with expressiveness - when mouthing the words demonstratively to indicate 'shouting'. All of these variations affect the coordinates of the facial landmarks in relation to each other, in an attempt to represent a wide array of situations in which the same type of grammatical facial expression is used, as well as how the same expression may not always look the same even on the same person. The closeness of the subject to the camera was constantly altered because the subject was leaning in and away from the stagnant camera to represent a typical conversation in which a person signing will move around. The placement of the face in the frame was also altered, the data captures the face in all areas of the frame of the picture. The face also tilted to the side/downward/forward to represent different camera angles. Each class of facial expressions has head tilts or nods that are customary to that particular expression and in that case that action was performed much more than the other head movements as it was more integral to the class. The subject was moving throughout the duration of the burst of photographs, as they were also signing various words and phrases pertaining to the class, to make the images more representative of those that would be analyzed in a translation or dictionary scenario.

The photograph bursts were taken at 24 frames per second in order to imitate the action of recording a video of a person signing and extracting individual frames. The images were square photos that are each 224 by 224 pixels. Each

Figure 2: Depiction of the 68 coordinates in the *dlib* facial landmark dataset.

photo was taken in color. The photographs were taken in front of a white background - painting canvas. The subject (Sasha Malysheva - researcher) was wearing a black shirt and her hair always began in the same way behind her back, the hair did move around somewhat when the subject was signing but it did not obscure the face with the exception of the ears. The computer from which the photos were taken was placed in the same spot for them all and did not move. The images were taken at the same time of day, and there was a light pointed at the subject and background from behind the computer, it was standing in the same position to make certain that the lighting and shadows were kept consistent from picture to picture.

It is clear that the above factors remaining constant heavily aid the inflate the success rate of the prediction models. In order for a model of similar nature to be remotely applicable in a real world scenario there would need to be a diverse population recording the data, but this project was limited in time and location.

5 Results and Discussion

5.1 Results

Results			
	F1 Score	Precision	Recall
Train	0.972738922403794	0.9727306765906399	0.9727476723571491
Test	0.961032819445758	0.9633330161448916	0.9606474933207606

5.2 Discussion

The F1 scores for this project are arbitrarily high and a direct result of the data used to train the model. The images used as the training dataset contained the same one person in each photo, therefore only one face was analyzed and the size of features and natural facial disparities between different were not a factor in this model. Furthermore every person has a certain way in which their face moves, there was not significant variation in the way these expressions were presented. There was also a consistent background and lighting. The background and lighting were kept stagnant in order to test whether this method of facial landmark tracking would work, so many factors were kept controlled. The only one subject was a culmination of the above reason as well as a limitation in access of other people who knew sign language who could record the data in the same format virtually. It is clear that these factors remaining constant heavily aid the inflate the success rate of the prediction models.

All the weight coefficients for the logistic regression of this model are relatively low, but some are substantially lower than others. In future works, it would be interesting to explore how many of these coordinates could be taken away, but still keep a reasonable F1 score. This could be one way of attempting to cut the execution time of the current model.

In order for a model of similar nature to be remotely applicable in a real world scenario there would first and foremost need to be a diverse population of people in the dataset. Subsequently, different backgrounds and lightings would need to be represented in the dataset If this method were to be applied as a translation tool or educational tool this model would need to be trained on a much more substantial dataset.

5.3 Future Work

There are several additions to this model that could be made that may aid in the efforts to create a standardized ASL dictionary. The first iteration would be to train this or a similar model on videos rather than photos to create a more realistic representation of ASL. Second, adding in a depth detection element - as some non manual markers have features of leaning in. Eventually, combining a grammatical facial recognition model together with a model that recognizes individual gestures in ASL. Then one that detects ASL grammar vs. English grammar - this could be a powerful educational tool for people unfamiliar with the ASL grammar system. Models such as this one could be used for live translation on the internet, particularly for ASL speakers who are illiterate or are unfamiliar with english grammar. Online educational tools for ASL could utilize models like this to examine the student's form and correct use of non-manual markers. Further exploration has the potential to unify the deaf and hearing community on a much greater scale. [18][19]

6 Conclusion

American Sign Language is a complex language with many facets that need to be considered when creating a translation model. ASL lacks a standardized dictionary and this strongly hinders its ability to be quickly translated and efficiently learned. Most current translation models focus heavily on the hands

and the specific signs that the subject is exhibiting and overlook non-manual markers such as facial expressions which are intrinsic to ASL's grammar. This project shows that in a controlled environment with sufficient data facial landmark tracking proves to be a very accurate metric by which to predict the class of grammatical facial expression. If this model is adapted for a diverse dataset and retains a relatively high F1 score then in the future this technology could be worked into a very powerful tool for American Sign Language translation, education, and an ASL stardardized dictionary.

References

[1]

[2]

[3]

[4]

[5]

[6] 2013.

[7] 2015.

[8] 2019.

[9] 2019.

[10] 2019.

[11] E B O N N I and Tucker. *Deaf Culture, Cochlear Implants, and Elective Disability*.

[12] Robert Bayley, Joseph Hill, Carolyn Mccaskill, and Ceil Lucas. *Attitudes towards Black American Sign Language*.

[13] Hope Dawson and Michael Phelan. *Language files: materials for an introduction to language and linguistics*. The Ohio State University Press, 2016.

[14] Tanya Denmark, Joanna Atkinson, Ruth Campbell, and John Swettenham. Signing with the face: Emotional expression in narrative production in deaf children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 49(1):294–306, Sep 2018.

[15] Rawini Dias. American sign language hand gesture recognition, Dec 2019.

[16] Eeva A. Elliott and Arthur M. Jacobs. Facial expressions, emotions, and sign languages. *Frontiers in Psychology*, 4, 2013.

[17] Michael Erard. Why sign-language gloves don't help deaf people, Nov 2017.

[18] Matt Huenerfauth and Vicki Hanson. Sign language in the interface: Access for deaf signers.

[19] Hernisa Kacorri, Pengfei Lu, and Matt Huenerfauth. *Evaluating Facial Expressions in American Sign Language Animations for Accessible Online Information*.

[20] Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. Real-time sign language gesture (word) recognition from video sequences using cnn and rnn. *Advances in Intelligent Systems and Computing*, page 623–632, 2018.

[21] David McNeill and Susan D Duncan. Grammar, gesture, and meaning in american sign language. *Sign Language Studies*, 5(4):506–523, 2005.

[22] H. E. Meador and P. Zazove. Health care interactions with deaf culture. *The Journal of the American Board of Family Medicine*, 18(3):218–222, May 2005.

[23] Nicholas Michael, Carol Neidle, and Dimitris Metaxas. *Computer-based recognition of facial expressions in ASL: From face tracking to linguistic interpretation.*

[24] Ross E Mitchell, Travas A Young, Bellamie Bachleda, and Michael A Karchmer. How many people use asl in the united states? why estimates need updating. *Sign Language Studies*, 6(3):306–335, 2006.

[25] opencv. opencv/opencv, Dec 2019.

[26] Adrian Rosebrock. Facial landmarks with dlib, opencv, and python - pyimagesearch, Nov 2018.

[27] Adrian Rosebrock. Face detection with opencv and deep learning - pyimagesearch, Feb 2019.

[28] Devesh Walawalkar. *Grammatical facial expression recognition using customized deep neural network architecture.*

[29] Frances Stead SellerscloseFrances Stead SellersSenior writer on the America desk EmailEmailBioBioFollowFollowSenior writer. Perspective — how america developed two sign languages — one white, one black.