Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 91–94
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Mouthing Recognition with OpenPose in Sign Language

María Del Carmen Sáenz
Computing & Digital Media, DePaul University
Chicago, IL. USA
msaenz@depaul.edu

Abstract

Many avatars focus on the hands and how they express sign language. However, sign language also uses mouth and face gestures to modify verbs, adjectives, or adverbs; these are known as non-manual components of the sign. To have a translation system that the Deaf community will accept, we need to include these non-manual signs. Just as machine learning is being used on generating hand signs, the work we are focusing on will be doing the same, but with mouthing and mouth gestures. We will be using data from The National Center for Sign Language and Gesture Resources. The data from the center are videos of native signers focusing on different areas of signer movement, gesturing, and mouthing, and are annotated specifically for mouthing studies. With this data, we will run a pre-trained Neural Network application called OpenPose. After running through OpenPose, further analysis of the data is conducted using a Random Forest Classifier. This research looks at how well an algorithm can be trained to spot certain mouthing points and output the mouth annotations with a high degree of accuracy. With this, the appropriate mouthing for animated signs can be easily applied to avatar technologies.

Keywords: Avatar technology, American Sign Language, OpenPose, Nonmanual signs, Mouthing, Mouth gestures

1. Introduction

Many Deaf people have American Sign Language (ASL) as their native language; their native tongue is usually secondary. Most people have limited reading and writing skills in said spoken language, leading to disadvantages in everyday situations such as health, education, and work. Communication barriers can occur especially in emergencies or government spaces. For example, if an emergency announcement is made on a train, there will be a delay in communication for a Deaf individual. An automatic translation system, such as an avatar, can provide rudimentary communication, in ASL on a public address system. These non-invasive technologies have been explored for the last 20 years to present sign languages. Many prototypes have been explored to accelerate Deaf-accessible systems, such as weather reports, airport security personnel, and government offices (Wolfe et. al, 2021).

Studies using a signing avatar combined with automatic translation systems, have focused on the hands more so than any other part of the avatar. Even though, it is well known that non-manual components of a sign, such as mouthing and mouth gestures, are used to discern signs that are closely related semantically as they may share the same movements or handshapes (Koller et. al, 2015). Mouthing itself is from spoken language in which you partially or fully mouth a word (Bickford and Fraychineaud, 2006). While mouth gestures come from the Deaf community, with no clear origin, such as mouthing "CHA" after signing the word "big" (Bickford and Fraychineaud, 2006). Just like in spoken language, the mood is conveyed with facial expressions and how words are said (mouthed). Having no facial expressions or mouthing/mouth gestures in sign language, according to Baldassarri et al., "is like speaking in a monotonic voice: more boring, less expressive and, in some cases, ambiguous" (2009).

Through the years as technology has advanced, so has avatar technology. However, there are still many inquiries regarding how to display information linguistically and pragmatically on the avatar's face (Wolfe et. al, 2021). Currently, work done with the face and mouth with present-day technologies available have long rendering times and can be incompatible with interactive graphic applications (Wolfe et. al, 2021).

Just like the work being done on algorithms for animating hand signs, this research aims to train how to spot mouthing points with exactitude to automate and apply it in avatar technologies for appropriate mouthing/mouth gestures for animated signs.

2. Related Work

The earliest research about mouthing, was in 1968 by Fisher (Koller et. al, 2015) distinguishing between a viseme and phonemes. Phonemes are the smallest units that compromise spoken language. While a viseme is made up of several speech sounds (phonemes). A viseme is "a set of phonemes which have an identical appearance on the lips" as they are the visual twin of phonemes (Bear and Harvey, 2017). As more research was being done in understanding how to visualize mouth movement to create speech, the audio-visual speech recognition field was born. This, in turn, led to the studying of the correlation of facial expression recognition with mouth shape creation via algorithms.

Usually mouthing and mouth gestures regarding sign language detection are overlooked (Koller et. al, 2015), but interest has been developing in this field (Antonakos et. al, 2015). Automatic Sign Language Recognition (ASLR) systems have been looking into the shape and motion of the mouth to determine critical cues versus ones done carelessly. For example, in ASL the tongue going through the front teeth is something done carelessly, therefore not a cue (Antonakos et. al, 2015). However, a critical cue is when one can recognize the state of the mouth. Such as open, closed, or very closed mouth during facial recognition (Koller et. al, 2015). Other related work has looked at using sequential pattern trees (Koller et. al, 2015) for general facial tracking or weak supervision models for facial features (Koller et. al, 2015). Overall, many models and analyses have been done on the face and head movements, which have partially included mouthing and/or mouth gestures.

On the other side, we must consider the progress in Computer Generated Imagery (CGI) and how it has advanced facial and mouthing in various spaces.

One of the first computer-animated faces called *Tony* from 1985, took 3 years to create a 7.5-minute film. Although it won many prizes for its innovation, this character today suffers from the phenomenon called *uncanny valley (*Wolfe et. al, 2021*)*; which gives the viewer a feeling of uneasiness or repulsion of seeing the humanoid figure. To avoid this, many animations use cartoon or alien humanoids, because since they are less human-like, they are more accepting of their emotions and expressions (Wolfe et. al, 2021). The best effects for facial and mouthing imagery in more complex visuals, still take hours to render a frame even though we have faster computers. There is also the painstaking task of doing some work manually, especially for frame transitions (Wolfe et. al, 2021).

With many advancements done in computational imagery and graphics, as well as in modeling, it takes time to fully capture the facial expressions. As well as creating reliable and most of all, believable mouthing, and mouth gestures. Just as many efforts are put into automating hand signs for signed languages, one must put in work on non-manual signs to have an avatar-based translation system be accepted in the Deaf community. The work proposed in this paper is attempting to bridge the gap in its usage of modeling and analyzing visual data to attempt to output mouthing points that can be used in automation for avatar usage.

3. Data Analysis

Motion capture is one way of data collecting to analyze sign languages. Much of this data is, again, primarily focused on the hands and how they move. Another way of studying sign languages is by using images or videos of native signers that are already available. OpenPose is a pre-trained Neural Network that analyzes video and images for a "real-time multi-person system to jointly detect human body, hand, facial, and foot keypoints." (Cao et. al, 2021). With 135 keypoints overall and 70 face keypoints, we will be analyzing videos of native signers which are publicly available.

3.1 Video Dataset

The dataset that is used was specifically captured to study ASL, which demonstrates the necessary parts of Sign Language accurately. The National Center for Sign Language and Gesture Resources (B.U., 1999), has a significant corpus of ASL videos of native signers. It contains multiple synchronized video files showing views from different angles and close-ups of the face. The corpus is a collection of 2,617 videos in MP4 format that has been compressed from 60 frames per second to 30 frames per second.

Figure 1: Example frames of video dataset

To coincide with each video, DePaul University has created an ELAN (also known as EUDICO annotation format) formatted file that groups different areas of the signer's mouthing and mouth gestures. The ELAN formatted file offered many mouthing annotations, but we focused on 9 annotations with a minimum of 35 examples as a requirement.

The 9 annotations we focused on were:

- Open and corners down
- Intense
- Raised upper lip
- Lips spread and corners down
- Lips pursed: mm
- Open (as in mouth open)
- Onset (mouth movement start)
- Offset (mouth movement end)

3.2 OpenPose Dataset

Although OpenPose has 70 face keypoint estimations that we can use on the video dataset, we will be focusing on points 48, 54, and 60-67 which pertain to the mouth.

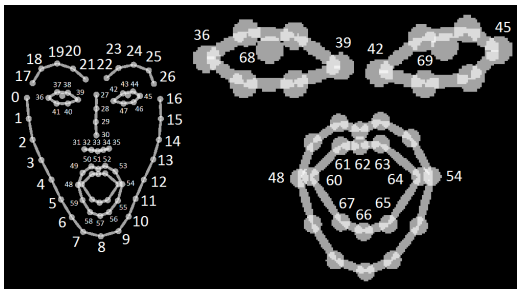

Figure 2: Facial keypoints in OpenPose

When we run the dataset through OpenPose the output visually shows the facial keypoints being mapped to the video.

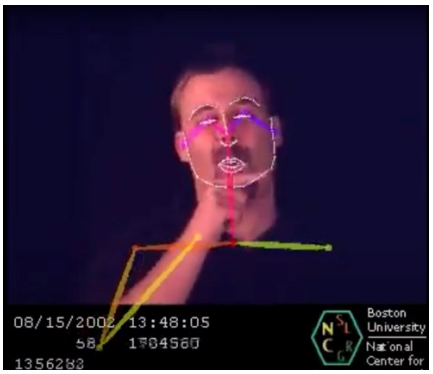

Figure 3: OpenPose keypoints on Video Dataset

4. Modeling

OpenPose is a powerful tool that was used to build highly confident mappings of the mouth. It works such that it uses two parallel divisions of convolutional network layers (Cao et. al, 2021); the first predicting 18 confidence maps, while the other predicts 38-part affinity fields. The confidence maps denote the specific part of the human pose skeleton, and the affinity fields denote the level of association between the parts (Cao et. al, 2021). In the last stages of the OpenPose algorithm, it cleans up its predictions made by the branches, weaker links are pruned via the PAF values, and the keypoints are then estimated and allocated on the video itself. Before OpenPose, some libraries were using different models such as Alpha-Pose and Mask R-CNN.

Comparing the runtime analysis of all 3, OpenPoses' runtime is constant, while Alpha-Pose and Mask R-CNN grow linearly with more people in the video. Although we are only focusing on one person in our video datasets, future work with multiple signers would be easier to evaluate using this software, especially with its constant runtime analysis.

After running OpenPose on 2,617 videos, we join the video JSON output with its respective ELAN annotation file by converting both into data frames and joining them via timestamp keyframe. This allowed us to analyze what annotations we wanted to focus on and at the same time have more than 35 videos available with said annotations. We were left with about 1,800 videos and used a matplotlib animator to manually look over the keyframes for occlusion and obstruction of the face by the hands. The filtering of the videos was only for extreme distortions and others were left to train the model effectively in the next phase.

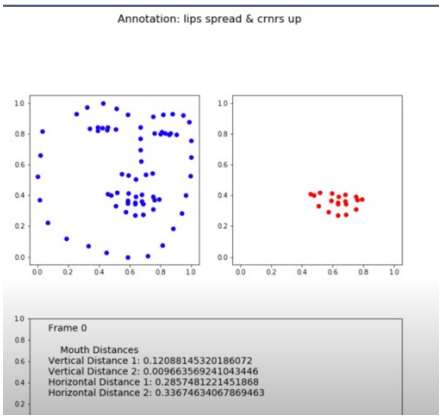

Figure 3: Animator used for looking over distortions

Combing through the data were left with 2,217 videos that had one or many of the annotations that we were interested in further analyzing using other modeling techniques. The next modeling technique we used, was a Random Forest Classifier (RFC) Model, an ensemble method, that has been utilized before to study Sign Languages (Su et. al, 2016). Going through the output of the OpenPose datasets, there was one sample size that had most of the data. To take advantage of this classifier, we used an oversampling method, called SMOTE (Synthetic Minority Oversampling Technique) (Chawla et. al, 2002), to improve the random oversampling. For comparison's sake, we ran the RFC without resampling and with resampling. A Grid Search was used to find the best hyperparameters for both the resampled and the non-resampled data, coming up with the same hyperparameters.

5. Results

Overall, the dataset showed a higher accuracy with the resampled data as opposed to the non-resampled data in the test balanced accuracy of the model and the validation accuracy of the annotations on the facial points themselves.

Dataset	Validation Accuracy	Test Balance Accuracy
With Resampling	0.96 (+/- 0.01)	0.6664373289281572
Without Resampling	0.43 (+- 0/03)	0.4392537365588655

Table 1: General Results of RFC

The classification reports also show that the recall is higher when there is more data to analyze for each facial keypoint and their respective annotation.

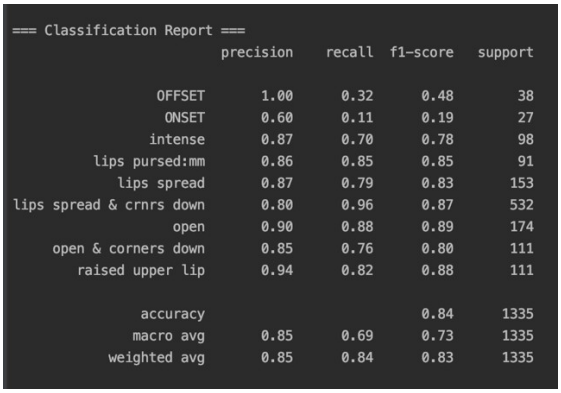

Figure 4: RFC without resampling of the dataset

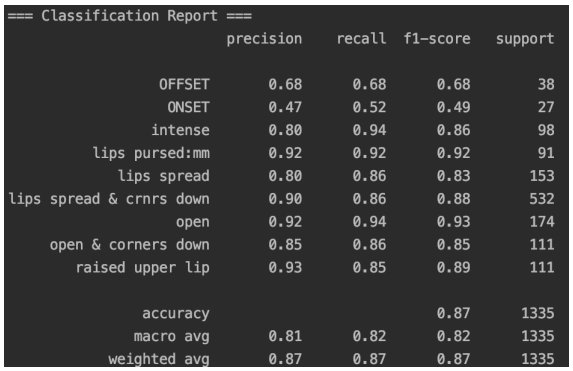

Figure 5: RFC with resampling of the dataset

6. Conclusion

A CNN with an RFC can prove to give a high accuracy in knowing which annotation is which on the facial keypoints. However, to have more balance in the tree, we need more data to work with from credible resources. Many institutions are sharing their corpus with other universities and agencies. Then we can add known annotations, like ELAN to the corpora that can assist in researching further the automation of mouthing and mouth gestures. Although the dataset used was small, we can see that a model can be trained to be effective in figuring out what mouth gestures are being used on specific facial points. For avatar translation systems, automation of the correct hand and mouthing/mouth gestures will be highly beneficial in getting us towards a system that will be acceptable to the Deaf community. As well as bridging the gap between the Hearing and Deaf communities.

7. Bibliographical References

Baldassarri, Sandra & Cerezo, Eva & Royo-Santas, Francisco. (2009). Automatic Translation System to Spanish Sign Language with a Virtual Interpreter. 5726. 196-199. 10.1007/978-3-642-03655-2_23.

Bear, Helen L. and Richard Harvey. "Phoneme-to-viseme mappings: the good, the bad, and the ugly." ArXiv abs/1805.02934 (2017): n. pag.

Bickford, J. and Fraychineaud, K. (2006). Mouth Morphemes in ASL: A closer look Theoretical Issues in Sign Language Research 9, 32–47.

Boston University. "Corpus of American Sign Language (ASL) Video Data from Native Signers". National Center for Sign Language and Gesture Resources at B.U., Dec. 1999, https://www.bu.edu/asllrp/cslgr/.

Chawla, N. et al. "SMOTE: Synthetic Minority Over-sampling Technique." J. Artif. Intell. Res. 16 (2002): 321-357.

E. Antonakos, A. Roussos and S. Zafeiriou, "A survey on mouth modeling and analysis for Sign Language recognition," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, pp. 1-7, doi: 10.1109/FG.2015.7163162.

Koller, Oscar & Ney, Hermann & Bowden, Richard. (2015). Deep Learning of Mouth Shapes for Sign Language. 10.1109/ICCVW.2015.69.

San-Segundo, Rubén & Montero, Juan & Cordoba, Ricardo & Sama, V. & Fernández-Martínez, Fernando & D'Haro, Luis & López-Ludeña, Verónica & Sánchez, D. & García, A.. (2012). Design, development, and field evaluation of a Spanish into sign language translation system. Pattern Analysis and Applications. 15. 10.1007/s10044-011-0243-9.

Su, Ruiliang & Xiang, Chen & Cao, Shuai & Zhang, Xu. (2016). Random Forest-Based Recognition of Isolated Sign Language Subwords Using Data from Accelerometers and Surface Electromyographic Sensors. Sensors. 16. 100. 10.3390/s16010100.

Wolfe, R., Hanke, T., Langer, G., Jahn, E., Worseck, S., Bleicken, J., McDonald, J. & Johnson, S. Exploring Localization for Mouthings in Sign Language Avatars. Proceedings of the International Conference on Language Resources and Evaluation (LREC) 2018

Wolfe, R., McDonald, J., Johnson, R., Moncrief, R., Alexander, A., Sturr, B., Klinghofer, S., Conneelly, F. Saenz, M. & Choudhry, S. State of the Art and Future Challenges of the Portrayal of Facial Nonmanual Signals by Signing Avatar. International Conference on Human- Computer Interaction pp. 639-655. Springer, Cham. 2021

Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.