An Improved Avatar for Automatic Mouth Gesture Recognition

Ronan Johnson
sjohn165@depaul.edu
DePaul University
Chicago, Illinois, USA

Maren Brumm
maren.brumm@uni-hamburg.de
University of Hamburg
Hamburg, Germany

Rosalee Wolfe
rwolfe@depaul.edu
DePaul University
Chicago, Illinois, USA

KEYWORDS

Sign Language, avatar technology, mouth gesture classification, neural networks

1 INTRODUCTION

This work is an extention of our paper submitted to the SignNonmanuals Workshop 2 [1]. Section 1 to 4 have been submitted there while Section 5 and 6 show new results.

Automatic sign language recognition has so far mainly focused on manual features. There have been some attempts to automatically classify mouthing [5] but to our knowledge there has been no prior work on automatic mouth gesture recognition. We aim to train an artificial neural network to classify mouth gestures from video. This helps the advancement of automatic sign language recognition and could be used to automate the annotation process of mouth gestures in large corpora such as the DGS-Korpus project. Neural networks require a large amount of training data which is excessively difficult to gather by manual annotation. To overcome this problem, we propose the use of avatar technology. The goal is to create a large number of animated video clips showing an avatar performing different mouth gestures, which will then be used as additional training data.

2 MOUTH GESTURE SELECTION

For the real world training and test data, we are using the corpus data of the DGS-Korpus project. We identified 21 mouth gestures that appear frequently. However, this list is by no means exhaustive. So far 3175 gestures have been annotated manually. But the number of occurrences of each gesture varies considerably. For instance, we found one gesture that had only 13 annotated examples. This highlights the difficulty of gathering sufficient training data by manual annotation, and that artificial data could be of great use.

3 REQUIREMENTS FOR AVATAR VIDEOS

To be useful for training the neural network, the avatar videos should mimic the real videos as much as possible. In addition to the avatar appearing natural and performing the mouth gesture correctly, there should also be a range of variations in the movements as can be found in the real videos. These variations include the intensity and the duration of the mouth gesture, the start and end pose of the head, and the movement of the head for the duration of the video. Ideally, there would also be variation in which avatar is used, but given there is only one currently available, this is impossible at the moment. Of course, the avatar is not able to mimic the gestures as naturally and in such variations as they appear in human conversations. Therefore, we will use exclusively natural videos in our testing data set. Because of this, the role of the synthetic data is currently ancillary to the recorded videos. Its true efficacy will be seen in comparing the performance of the classifier when trained on the natural videos vs. the generated animations.

4 CREATION OF AVATAR VIDEOS

In order to create the number of animations necessary for sufficient training data, we developed a script that randomizes all the necessary parameters for creating the amount of variation needed. We determined an appropriate range of values for the head positions and movements by the most extreme poses that were likely to occur within the real video reference. The script is fully automated such that the user need only set the mouth gesture to be replicated and the number of desired animations. It then proceeds to generate and render each animation individually. This ameliorates the difficulty of gathering huge quantities of annotated training data.

5 ISSUES WITH THE AVATAR REPRESENTATION

To create the mouth shapes, the avatar Paula was rigged using the MPEG-4 hAnim mouth landmarks, which defined an outer and inner lip contour (Figure 1) [4]. The locations of the landmarks controlled the deformation of the avatar's mouth. However, this rig did not facilitate the creation of all of the DGS mouth gestures. Unfortunately, simply using these as control points in the rig these did not contain enough expressiveness to generate all mouth gestures. To identify the problems, we used Ekman's list of action units (AUs) of contraction or relaxations of facial muscles as a guide [3]. Most of the problems arose in representing the effects of the orbicularis oris muscle that encircles the mouth. We found that the action unit of Lip Funneler (AU 22) was not possible with the current rig, but it is necessary to express DGS mouth gestures. This action unit pushes the lips outward as if saying the word "flew". Another essential action unit is the Lip Suck (Ekman's AU 28), where the lips curl inward over the teeth. This action is present in at least three of the DGS mouth gestures.

We created a new rig, which increased the number of controls surrounding the outer mouth from 10 to 44. These create more control over shaping the outer mouth perimeter. To these, we attached an additional set of controls to shape the lip profile. In Figure 3 the new controls appear as small spikes. The base of each spike is located on the perimeter of the outer mouth. The spike protrudes into the lip volume. To produce the appearance of a Lip Suck, the animator rotates the spikes toward the interior of the mouth. To product the appear of the Lip Funneler, the animator rotates the

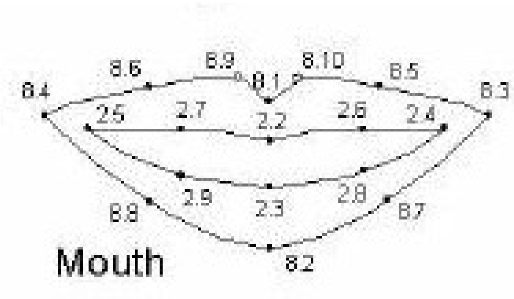

Figure 1: MPEG-4 hAnim mouth landmarks

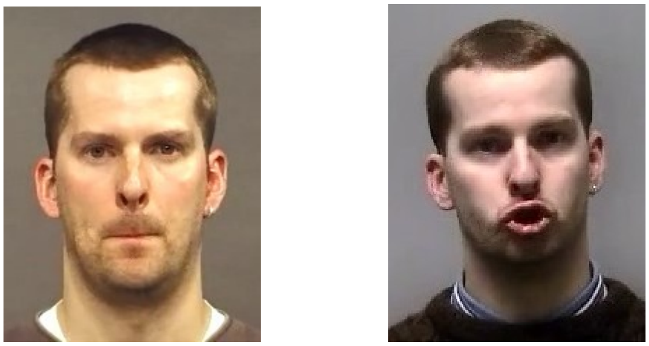

Figure 2: Mouth gestures impossible to produce with current avatar

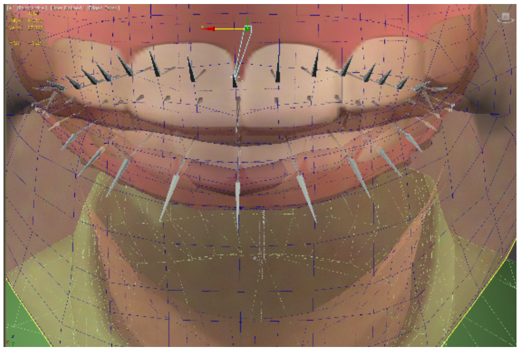

Figure 3: New mouth rig

spikes outward.

Figure 4 demonstrates the improvement in the appearance. The left image is an example of a gesture that uses Lip Suck. The image in the middle is the closest approximation that was possible with the previous mouth rig. The image on the right is from a version of the avatar that uses the improved rig.

6 FIRST CLASSIFICATION RESULTS

To classify the different mouth gestures we use a spatiotemporal convolutional, residual network combined with bidirectional recurrent units based on [6] but use the visual part of the network only. The network was pretrained on the Lip Reading in the Wild database [2].

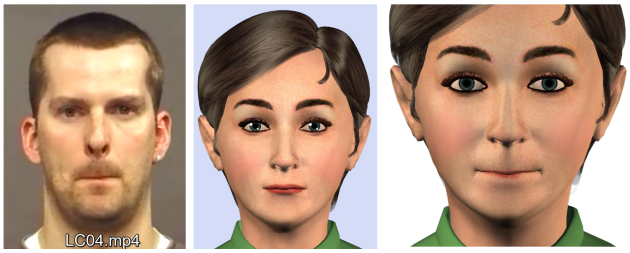

Figure 4: Lip Suck, Old rig, new rig

As a first test we trained the classifier to differentiate between five mouth gestures. One experiment was conducted with natural training data only, one additionally included 1000 avatar videos per class. Due to timing these were made with the old rig. As test data we use natural data only. We use 10-fold cross-validation to make use of all natural data.

When including the avatar data in the training set a peak accuracy of 84.87% is reached after 17 epochs, which is a slight improvement over the results without avatar videos. Here a peak accuracy of 83.59% is reached after 13 epochs.

However, these are very first results only. We expect the avatar videos to have more of an impact when more mouth gesture classes are involved and the new mouth rig to further improve the classification results.

REFERENCES

[1] Brumm, M., Johnson, R., Hanke, T., Grigat, R.-R., and Wolfe, R. Use of avatar technology for automatic mouth gesture recognition. Poster presented at Sign-Nonmanuals Workshop 2, May 3-4 , 2019 at University of Graz, Austria, 04 2019.

[2] Chung, J. S., and Zisserman, A. Lip reading in the wild. In *Asian Conference on Computer Vision* (2016), Springer, pp. 87–103.

[3] Cohn, J. F., Ambadar, Z., and Ekman, P. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment* (2007), 203–221.

[4] Fratarcangeli, M., and Schaerf, M. Realistic modeling of animatable faces in mpeg-4. In *Computer Animation and Social Agents* (2004), pp. 285–297.

[5] Koller, O., Ney, H., and Bowden, R. Deep learning of mouth shapes for sign language. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2015), pp. 85–91.

[6] Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., and Pantic, M. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), IEEE, pp. 6548–6552.