Use of Avatar Technology for Automatic Mouth Gesture Recognition

Maren Brumm[1], Ronan Johnson[2], Thomas Hanke[1], Rolf-Rainer Grigat[3], Rosalee Wolfe[2]
[1]Institute of German Sign Language and Communication of the Deaf, University of Hamburg, Gorch-Fock-Wall 7, 20354 Hamburg, Germany
[2]School of Computing, DePaul University 243 S. Wabash Chicago IL 60604 USA
[3]Vision Systems, Hamburg University of Technology, Harburger Schloßstraße 20, 21079 Hamburg, Germany
maren.brumm@uni-hamburg.de, sjohn165@depaul.edu, thomas.hanke@uni-hamburg.de, grigat@tuhh.de, rwolfe@depaul.edu

1 Aim of the Project

Automatic sign language recognition has so far mainly focused on manual features. There have been some attempts to automatically classify mouthing [2] 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 Training the Classifier

We use a spatiotemporal convolutional, residual network combined with bidirectional recurrent units based on [3] but use the visual part of the network only and pretrain it on the Lip Reading in the Wild database [1]. So far we have classification results only for the ten mouth gestures appearing most frequently out of all 21 mouth gestures. First tests show that we can reach a classification accuracy of 68%. We plan to use the artificial data either for further pretraining or directly as additional training data. By doing so we hope to improve our results and make classification possible for all of the 21 mouth gestures, even those with very few real world examples.

References

[1] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[2] Oscar Koller, Hermann Ney, and Richard Bowden. Deep learning of mouth shapes for sign language. In *Deep Learning of Mouth Shapes for Sign Language*, pages 477–483, 2015.

[3] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552. IEEE, 2018.