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]University of Hamburg Institute of German Sign Language and Communication of the Deaf Hamburg, Germany
[2]DePaul University School of Computing Chicago, USA
[3]Technical University of Hamburg, Germany

Aim of the Project

Automatically classifying mouthing has been attempted, but automatically classifying mouth gesture recognition has not.

We aim to train an artificial neural network to classify mouth gestures from video, which can be used to automate the annotation of large corpora such as the DGS-Korpus Project.

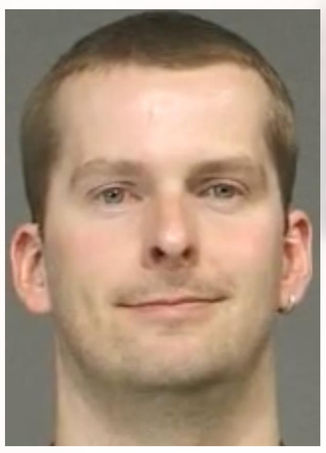
We propose using avatar technology to generate additional training data needed to increase the accuracy of the classifier.

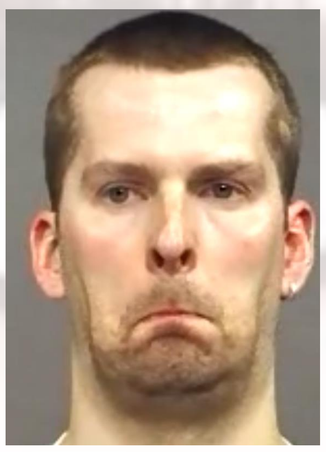
The goal is **to create a large number of animated video clips** showing an avatar performing different mouth gestures.

Mouth Gesture Selection

For the real world training and test data, we are using the corpus data of the DGS-Korpus project.

- 3175 gestures manually annotated so far in the corpus, with 21 frequently-occurring gestures identified.
- Some of these gestures have very few occurrences, as few as 13.





Requirements for Avatar Videos

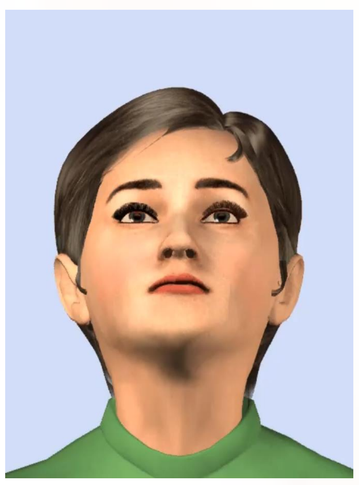
To be useful for training the neural network, **the avatar videos should mimic the real videos as much as possible,** through accuracy and naturality.

There must also be a wide amount of variation between the videos, such as:

- Intensity and duration of the mouth gesture.
- Start and ending pose of the head.
- Speed and direction of the head during the video.

Because the avatar cannot perfectly mimic natural video, this synthetic data is currently ancillary to the recorded videos.

Variation 1



Start

End

Variation 2

Start

End

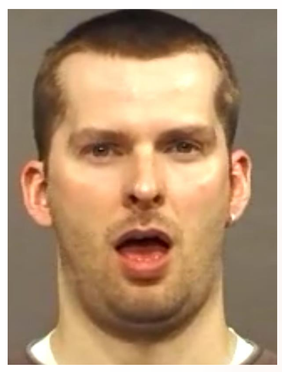
Creation of Avatar Videos

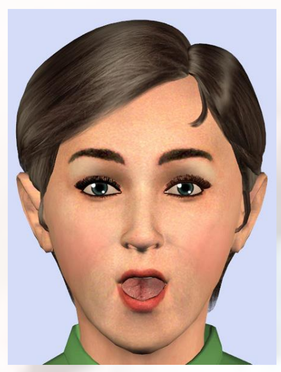
We utilized a Python script in order to generate the videos procedurally.

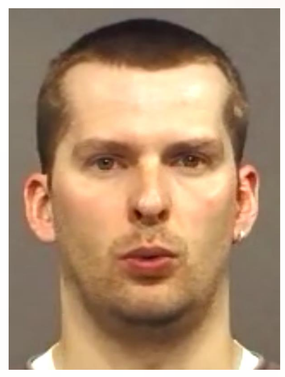
- Range of values is determined by the most extreme poses found in the video reference.
- 1000 animations generated per gesture.

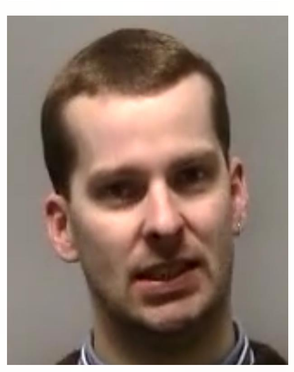
Automation allows for a huge amount of training data to be generated in minimal time.

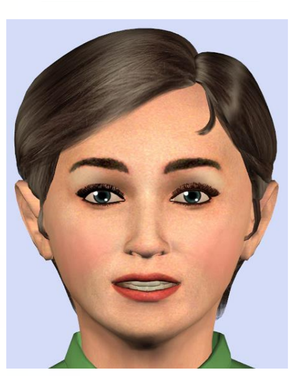
Gestures generated so far:











Training the Classifier

We use a spatiotemporal convolutional, residual network combined with bidirectional recurrent units.

We currently 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%.**

Future Work

We plan to use the artificial data either for **further pretraining or directly as additional training data.**

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

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

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.

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.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DGS-KORPUS

DePaul University

Akademie der Wissenschaften in Hamburg