Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 27–32
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Optimised Preprocessing for Automatic Mouth Gesture Classification

Maren Brumm[1, 2], Rolf-Rainer Grigat[2]
[1]Institute of German Sign Language and Communication of the Deaf,
University of Hamburg, Gorch-Fock-Wall 7, 20354 Hamburg, Germany,
[2]Vision Systems, Hamburg University of Technology, Harburger Schloßstraße 20, 21079 Hamburg, Germany
maren.brumm@uni-hamburg.de, grigat@tuhh.de

Abstract

Mouth gestures are facial expressions in sign language, that do not refer to lip patterns of a spoken language. Research on this topic has been limited so far. The aim of this work is to automatically classify mouth gestures from video material by training a neural network. This could render time-consuming manual annotation unnecessary and help advance the field of automatic sign language translation. However, it is a challenging task due to the little data available as training material and the similarity of different mouth gesture classes. In this paper we focus on the preprocessing of the data, such as finding the area of the face important for mouth gesture recognition. Furthermore we analyse the duration of mouth gestures and determine the optimal length of video clips for classification. Our experiments show, that this can improve the classification results significantly and helps to reach a near human accuracy.

Keywords: Sign Language Recognition/Generation, Machine Translation, SpeechToSpeech Translation, Statistical and Machine Learning Methods

1. Introduction

Mouth gestures are facial expressions in the context of sign language, that do not refer to words of a spoken language. They are an important part of the German Sign Language that can be crucial for understanding the meaning of signing (Von Agris et al., 2008).
A corpus with annotated mouth gestures would be helpful for further research, but is very time-consuming to acquire. The aim of this work is to develop a method to automatically classify mouth gestures by training a neural network. This could eliminate time-consuming manual annotations as well as advance automatic sign language translation. However, mouth gesture classification is a challenging task even for humans. Some mouth gesture classes are very similar to each other and the style, duration and intensity of a mouth gesture can vary significantly from person to person. Another issue is the small size of training material available. Therefore, careful preprocessing of the data can significantly improve the results, as it helps to reduce the input data to the necessary information only. We evaluate the effect of the usage of different regions of interest (ROIs) within a frame, different methods to convert the videos to a fixed length, as well as different clip durations.
Earlier works on non-manuals use facial landmarks as features. (Neidle et al., 2014) detect non-manual grammatical markers and (Luzardo et al., 2014) estimate a mouth state (open / close / narrow /...) by geometric features based on facial landmarks.
More recent works often use neural networks. To our knowledge there are only two publications on automatic mouth gesture recognition (Wilson et al., 2019), (Brumm et al., 2019). We extend the work of (Brumm et al., 2019), however, without the use of an avatar. Our work is also similar to (Wilson et al., 2019) but we use a different neural network architecture.
There are some papers on the related subject of mouthing, which are facial expressions in sign language that do refer to spoken words, such as (Koller et al., 2014), (Koller et al., 2015) and more on the broader field of spoken word recognition and lip reading like (Chung and Zisserman, 2016), (Chung et al., 2017), (Afouras et al., 2018) and (Martinez et al., 2020). The architecture of the neural network used in this work is based on (Petridis et al., 2018), who use spatiotemporal convolution followed by a 34-layer ResNet and 2-layer BGRU.

2. Dataset

Our dataset was generated from the DGS corpus of the DGS-Korpus project at the University of Hamburg [1]. It consists of 4177 mouth gestures from 281 different signers appearing in natural conversation. We identified 21 classes of mouth gestures, that appear frequently in the corpus. They were annotated independently by two different annotators. The annotators also provided the exact start and end point of the mouth gestures within the video.
However, for some of the 21 mouth gesture classes we could not find a sufficient number of training examples and in some cases two of the mouth gesture classes are too similar to be differentiated with a reasonable accuracy even for human annotators. We therefore reduced the number of mouth gesture classes for automatic classification to ten, by combining very similar mouth gestures to one mouth gesture class and leaving out classes with less than 52 examples.
This results in a dataset with 2842 examples of ten different mouth gesture classes. The number of examples per class varies between 52 and 615. Table 1 describes the ten chosen classes and shows how many examples are in the dataset for each of them.
To estimate the accuracy with which humans can perform mouth gesture classification, we use the inter-annotator agreement of the two annotators. As the annotators were originally asked to classify the data to 21 different mouth

[1] https://dgs-korpus.de

gesture classes, we can not determine the exact human classification accuracy (or inter-annotator agreement) for our ten class classification problem. Considering only examples where both annotators give a class within the ten chosen classes, the accuracy is 79.13%. Considering all clips where the first annotator gives a class within the ten chosen classes the accuracy is 66.40%. The real human accuracy is somewhere in this range.

3. Neural Network Architecture

The architecture of the neural network used is based on the work of (Petridis et al., 2018). However, we only use the visual stream of their two-stream network. It consists of a spatiotemporal convolutional layer, a 34-layer ResNet and a 2-layer BGRU. The network was pretrained on the Lip Reading in the Wild (LRW) database (Chung and Zisserman, 2016).

4. Proposed Preprocessing Options

4.1. Region of Interest

The original videos show the upper body of the person as well as the background, as can be seen in figure 1a. The first step is therefore to extract the region of interest (ROI). Our aim is to make the ROI as small as possible without loosing relevant features. This is especially important as our dataset is small, which makes it more difficult for the network to distinguish between relevant and non-relevant artefacts in the image.

We consider three possible ROIs shown in figures 1c, 1d and 1e. The first is a close-up of the mouth. The second shows the lower part of the face, as helpful features may also be located on the cheeks or the nose. The third option is to use the whole face, to also include possible features located on the eyes, eyebrows and forehead.

Figure 1b shows the ROI that was used in (Petridis et al., 2018). As we use their pretrained model, similarity effects have to be taken into account, as described in section 5.3.

4.2. Frame Sampling

Naturally the mouth gestures differ in length. We consider two different methods to transform the mouth gesture clips to a fixed length.

The first is to up- or downsample the clips to the required number of frames, as described in (Wilson et al., 2019). If a clip is too long, frames are removed at even intervals. If it is too small, frames are doubled at even intervals. This assures, that the mouth gesture is visible from the start until the end. But frames in between might be missing or doubled.

The second option we propose, is to cut out a consecutive number of frames left and right of the midpoint of the mouth gesture. This may lead to parts of the mouth gesture being cut off, if the actual mouth gesture is longer than the number of frames used or other video material being included that is not part of the mouth gesture, if the mouth gesture is shorter. But the clip that is cut out is consecutive. Both methods require knowledge of the location of the mouth gesture in the video. However, this information is available during training only. When applying the method to unlabelled data the location of the mouth gesture can only be approximated by the location of the hand gesture accompanying the mouth gesture. In this case the continuous method might be advantageous as the network has learned to classify clips that are incomplete or show unrelated material.

Section 6.2 shows the comparison of the different sampling methods. The results on approximating the mouth gesture location are shown in section 6.3.

4.3. Clip Duration

The duration of the clips used as input to the neural network is an important parameter. If a clip is too short, parts of the mouth gesture are cut off. If it is too long, it shows facial actions not related to the mouth gesture. Both aspects could lead to poorer training results. This is especially true if the exact timing of the mouth gesture is unknown and the starting and end point is determined by the hand gesture that is accompanying the mouth gesture, which would be the case in a real world scenario.

An analysis of the distribution of the length of mouth gestures can be seen in figure 2. It shows the box plot of the distribution. The length is given in number of frames, where all videos have been recorded at 50 frames per second. It can be seen that the majority of mouth gestures are relatively short. The mean is 24.9 frames and the 75th percentile 31 frames. Nevertheless the length can vary substantially. 5.1% of the mouth gestures have more than 60 frames and outliers reach up to 224 frames.

In our experiments we test a range of durations from 19 to 45 frames.

5. Experiments

5.1. General Preprocessing

We use OpenPose (Cao et al., 2018), (Simon et al., 2017) to detect facial landmarks on the face. These are used to transform the image so that the distance between the eyes is the same for all frames and all persons. We normalise the scale of the frames by the interocular distance and rotate them so that the axis between the eyes is horizontal. After alignment the ROI is extracted as described in 4.1.

The video clips are converted from RGB to grayscale, as previous tests showed no significant difference in the results. All frames are scaled to 96×96 pixels and normalised with the overall mean and standard deviation of the dataset.

As the number of examples per class differ a lot, the dataset is balanced by over- or undersampling classes to the median of examples per category.

5.2. Training

We use the pretrained model for the visual stream of (Petridis et al., 2018) and train the network end-to-end. The initial learningrate is set to $3 \cdot 10^{-4}$ in the ROI experiments and $3 \cdot 10^{-5}$ in the frame sampling and clip duration experiments, as the latter proved to be better in intermediate tests.

For data augmentation the data is cropped randomly to 88×88 pixels and randomly flipped horizontally during training.

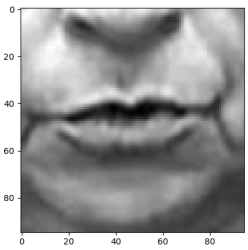
As our dataset is small, we use 10-fold cross validation

mouth gesture	description	number of examples
MO04	Lips open and stretched, teeth together and visible. Like german 'sss' or 'pss'.	98
MO07/LR03	Lips round and open. Like german 'o' .	167
MO08	Mouth wide open. Like german 'a' .	113
LR01	Lips round and puckered, air streams out through small opening. Like german 'sch'.	52
LR02/LR10	Lips pursed.	420
LR05/CH01	Blow out air continuously through rounded lips, cheeks possibly puffed.	615
LC04/LC05	Lips closed and stretched strongly, lips possibly sucked in.	556
LC06	Lips closed, corners of mouth curved down, lips possibly sucked in.	340
TO01/TO04	Mouth open, tongue protrudes or dorsum pressed to front.	264
TE03	Mouth slightly open, upper teeth on lower lip, sudden release of air. Like german 'pf'.	216

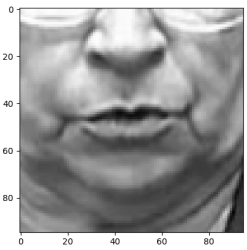
Table 1: Description of the ten mouth gesture classes used for classification and the number of annotated examples per class.

(a) original frame

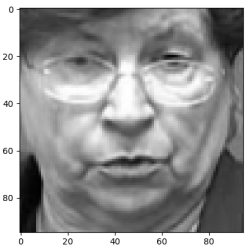
(b) pretraining ROI



(c) mouth ROI



(d) lower face ROI



(e) full face ROI

Figure 1: Example of an original frame (a) and our proposed ROIs (c)-(e). (b) shows the ROI used in (Petridis et al., 2018), with which our network was pretrained.

to make use of all data in the training- as well as in the validation set. This makes the results also more stable to statistical variations in the training procedure as all results are the combination of 10 individual training runs. Due to the dataset size we do not use a testset. All given results are the peak accuracy on the combined validation sets.

5.3. Experiments on ROI

Clips are cut to 29 frames using the up- and downsampling method described in 4.2.
When using a pretrained model one might achieve better results with inputs similar to the previous training material. Our network was pretrained using a ROI that is similar to our 'mouth only' ROI, see Figure 1. To ensure that we choose the best ROI for our dataset and not simply the one closest to the pretraining data, we run our experiment twice. The first time we use the pretrained model, the second time we train the network from scratch, to avoid influence from pretraining.

5.4. Experiments on Frame Sampling

For the experiments on the frame sampling methods we use the lower face ROI and cut the videos to 29 frames using either sampling method. We use the same sampling method for training and validation set.

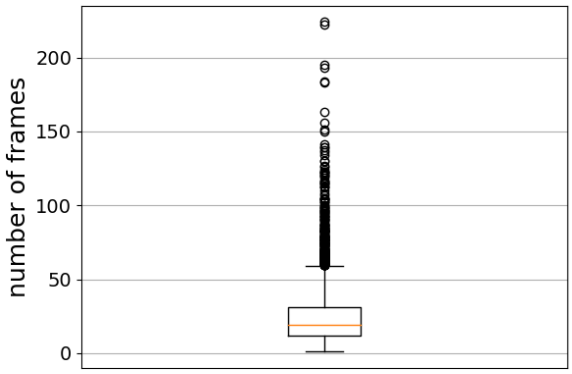


Figure 2: Box plot of the number of frames per mouth gesture for a frame rate of 50.

5.5. Experiments on Clip Duration

We use the' lower face' ROI and consider a number of frames from 19 to 45. For training, the videos are cut using the exact timing information given by the annotators. We cut the videos so that the midpoint matches with the midpoint of the true location of the mouth gesture. Therefore the clips may show more or less than the mouth gesture, but the mouth gesture is centred within the clip.
For the validation set we use two different options. One is to center the clip at the midpoint of the mouth gesture, as done with the training set. However, this information is not available in a real world scenario. Here we can only use the midpoint of the hand gesture as an approximation of the midpoint of the mouth gesture. This change may have an influence on the optimal number of frames, as more frames might be needed, to ensure that enough of the mouth gesture is included, if the timing of the mouth and hand gesture differ significantly. Therefore we validate the training results with both cuts.
Due to timing issues, the dataset was updated during the experiments. We therefore run the experiment with 29 frames twice, once on the old and once on the new version of the dataset to make the results comparable.
As the dataset used for pretraining consists of clips with a length of 29 frames, this might influence the results. We therefore rerun part of the experiment with the network trained from scratch.

6. Results

6.1. ROI

Table 2 shows the results for the three different ROIs with and without pretraining. For all ROIs the pretrained results are clearly better. In both the untrained and pretrained case the 'full face' ROI results in the lowest accuracy. So if there are helpful features on the upper part of the face, they are not strong enough to compensate the lower resolution of the images and the inclusion of unnecessary artefacts such as hair.
Without pretraining the 'mouth only' ROI gives better results than the 'lower face' ROI. With pretraining the 'lower face' ROI is better. This is a surprising result as the ROI used for pretraining is more similar to the 'mouth only' ROI. Therefore, the cause for the better results of the 'lower face' ROI in the pretrained case can not be, that the inputs are more similar to the pretraining inputs. Instead, the reason might be that a larger ROI is more complex to analyse. So the untrained network might fail to find the right features here and prefer more focused images, while the pretrained network already learned to find these features with the help of a much larger dataset and therefore benefits from the larger ROI with more features.
Therefore the 'lower face' ROI is preferable when using the pretrained network.

ROI	without pretraining	with pretraining
whole face	58.18	66.76
lower face	60.56	**70.60**
mouth only	**62.08**	68.93

Table 2: Peak accuracy for different ROIs.

6.2. Frame Sampling

The results for the frame sampling methods can be found in Table 3. The up- and downsampling method reaches a peak accuracy of 69.89 %, the continuous method 70.28 %. So the results for the continuous method are slightly better, but there is no significant difference. It seems, that it is at least equally important, that the clips are consecutive, to that the mouth gesture is cut exactly from start until end. The reason for that might be that the spatiotemporal convolution works best for consecutive frames. If several frames are doubled the layer can not extract any temporal information. If too many frames are deleted the facial movements might be too large and important frames might be skipped.

sampling method	accuracy
up-/downsampling	69.89
continuous	**70.28**

Table 3: Peak accuracy for different sampling methods.

6.3. Clip Duration

Table 4 shows the results for different clip durations, ranging from 19 to 45 frames with a frame rate of 50 frames per second. Here the pretrained model was used as starting point. As described in section 5.5, we used two different versions of the dataset and ran the experiment twice with 29 frames to make the results comparable. The change of dataset is indicated by the dashed line in the table.
When the videos are cut using the hand gesture position the accuracy decreases as expected. It is on average 2.7 percentage points lower. Apart from that, the results for both cuts are very similar.
Using less than 29 frames clearly worsens the results. For

29 to 39 frames there is no significant difference in the achieved accuracy. For 45 frames the accuracy drops for both cuts. We assume that for more than 39 frames the clips involve too much other material, that is not part of the mouth gesture.

Table 5 shows the classification accuracy for different numbers of frames, when the network is trained from scratch. Again, the results for the hand gesture cut are less accurate than for the mouth gesture cut, but apart from that the results are similar. However, when the network is trained from scratch shorter clips are clearly preferable. Here 19 frames achieve a better result than 29 or 39 frames. The reason might be that the pretrained network prefers 29 frames because that is what it was pretrained on. However, an argument against that is, that 35 and 39 frames achieve similar results in the pretrained case. Another, possibly additional, reason might be, that the untrained network prefers 19 frames because that is less data to process and features are easier to spot, which is not such a problem for the pretrained network. To find out which is the case here, it would be necessary to cut the videos of the LRW dataset to less frames, train the network with it and use this as a pretrained model for further experiments. However, this is beyond the scope of this work. Another possibility would be to create a model that is less biased to the number of frames in the clips, by performing variable length augmentation as described in (Martinez et al., 2020). In this case it might also be beneficial to use the mouth gesture data with its actual varying length. For the training data the exact length is known, for application on unlabelled data, it might be possible to estimate it by the hand gesture length, if available.

number of frames	mouth gesture cut	hand gesture cut
19	68.07	65.29
25	69.14	66.49
29	70.28	67.16
29	70.47	**68.04**
35	70.40	67.76
39	**70.78**	67.94
45	69.55	67.33

Table 4: Peak accuracy for different number of frames, using the pretrained model.

number of frames	mouth gesture cut	hand gesture cut
19	**58.03**	**55.82**
29	55.25	53.07
39	53.45	51.02

Table 5: Peak accuracy for different numbers of frames, training from scratch.

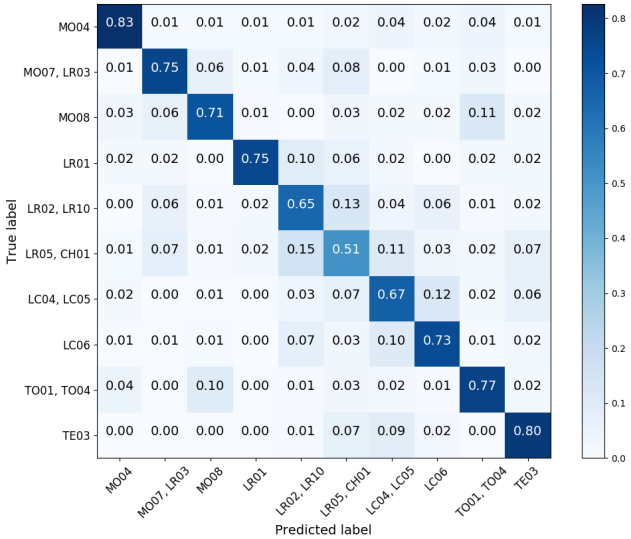


Figure 3: Confusion matrix for classification with continuous 29 frames centred at the hand gesture position.

6.4. Overall Results

Combining the results from all experiments, the best results are achieved with the pretrained model when using the 'lower face' ROI together with the continuous sampling method and a duration of 29 frames, which is equal to 0.58 seconds. For this setting we achieve an accuracy of 70.47% using the mouth gesture cut and 68.04% using the hand gesture cut, which would be used in a real world scenario. These results are comparable to human accuracy, which is between 66.47% and 79.13% for the given dataset.

Figure 3 shows the confusion matrix for the latter setting. It can be seen that the per class accuracy varies substantially from class to class, as some classes are well-defined while others overlap. For example, the round lips in class LR02/LR10 are similar to the lip shape when blowing air, as in LR05/CH01. The vibrating lip pattern of MO04 on the other hand, is unique and therefore easier to distinguish.

Interestingly, the number of examples per class in the dataset does not seem to have a high impact on the per class accuracy, as MO04 has the second least number of examples but the highest accuracy, while the three classes with most examples have the lowest accuracy.

7. Conclusion

In this work we compare different preprocessing options for mouth gesture classification from video.

The experiments on using different ROIs show that the best results can be achieved with a ROI that shows the lower half of the face. We compared two methods to format the videos to a fixed length: up- or downsampling the frames, so that the mouth gesture is shown exactly from start until end or using a time window of continuous frames centred to the middle of the mouth gesture duration. Both show similar results. We favour the continuous method, as it requires less information. Another important parameter is the duration of the videos, to make sure the relevant parts of the mouth gesture are included, but not too much additional material. We tested a range of 19 to 45 frames and showed that a

length of 29 frames is best, when using the pretrained network. When training from scratch, less frames are preferable.

Combining the results of all our experiments we achieve the highest accuracy when using the 'lower face' ROI and a duration of 29 continuous frames. If the clips are centred with the information of the exact mouth gesture we achieve an accuracy of 70.47%. If we use this information for training only and not for testing, as would be the real world scenario, the accuracy is 68.04%. In both cases we achieve results comparable to human accuracy, which lies in between 66.47% and 79.13%.

8. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

9. Bibliographical References

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Brumm, M., Johnson, R., Hanke, T., Grigat, R.-R., and Wolfe, R. (2019). Use of avatar technology for automatic mouth gesture recognition. In *SignNonmanuals Workshop 2, Graz, Austria, 2019*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *Asian Conference on Computer Vision*.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Koller, O., Ney, H., and Bowden, R. (2014). Read my lips: Continuous signer independent weakly supervised viseme recognition. In *European Conference on Computer Vision*, pages 281–296. Springer.

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

Luzardo, M., Viitaniemi, V., Karppa, M., Laaksonen, J., and Jantunen, T. (2014). Estimating head pose and state of facial elements for sign language video. In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, Language Resources and Evaluation Conference*.

Martinez, B., Ma, P., Petridis, S., and Pantic, M. (2020). Lipreading using temporal convolutional networks. *arXiv preprint arXiv:2001.08702*.

Neidle, C., Liu, J., Liu, B., Peng, X., Vogler, C., and Metaxas, D. (2014). Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in american sign language (ASL). In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, Language Resources and Evaluation Conference*.

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

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153.

Von Agris, U., Knorr, M., and Kraiss, K.-F. (2008). The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE.

Wilson, N., Brumm, M., and Grigat, R.-R. (2019). Classification of mouth gestures in german sign language using 3d convolutional neural networks. *International Conference on Pattern Recognition Systems ICPRS 2019*.