ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Disambiguating Signs: Deep Learning-based Gloss-level Classification for German Sign Language by Utilizing Mouth Actions

Dinh Nam Pham, Vera Czehmann and Eleftherios Avramidis

German Research Center for Artificial Intelligence (DFKI)
Speech and Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany

Abstract. Despite the importance of mouth actions in Sign Languages, previous work on Automatic Sign Language Recognition (ASLR) has limited use of the mouth area. Disambiguation of homonyms is one of the functions of mouth actions, making them essential for tasks involving ambiguous hand signs. To measure their importance for ASLR, we trained a classifier to recognize ambiguous hand signs. We compared three models which use the upper body/hands area, the mouth, and both combined as input. We found that the addition of the mouth area in the model resulted in the best accuracy, giving an improvement of 7.2% and 4.7% on the validation and test set, while allowing disambiguation of the hand signs for most of the cases. In cases where the disambiguation failed, it was observed that the signers in the video samples occasionally didn't perform mouthings. In a few cases, the mouthing was enough to achieve full disambiguation of the signs. We conclude that further investigation on the modelling of the mouth region can be beneficial of future ASLR systems.

1 Introduction

Sign Languages (SLs) are visual natural languages primarily used by deaf and hard of hearing people. Automatic Sign Language Translation (ASLT) could help the communication between signers and non-signers. One of the research tasks that may function as a building block for ASLT is the Automatic Sign Language Recognition (ASLR) which focuses on the recognition of isolated signs. Since SLs are multi-channel visual languages, ASLR is a challenging task, still in its infancy compared to Automatic Speech Recognition [1]. Although ASLR has had increasing attention, it is striking that most research mainly focuses on the hands to recognize signs. Despite non-manual cues being considered important elements of sign languages in linguistic studies e.g. on the German SL (DGS), only 8% of ASLR results from 2015 until 2020 employed them [2].

Particularly the mouth region should be of great interest since mouth actions have at least 3 functions: meaning specification, sole carrier of meaning and disambiguation [3]. The latter means that mouth actions can be used to distinguish homonyms (like Schwester/Bruder -sister/brother- in DGS), which have the same hand form and movement but different mouth actions. This ambiguity of signs poses a challenge for developing accurate and reliable ASLR systems.

We aim at measuring the importance of mouth actions for ASLR by exploring the performance of a classifier that recognizes ambiguous isolated signs in DGS.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

We train a deep learning model and investigate the impact of using different visual inputs on the performance of the model, including (a) the upper body with hands, (b) the mouth and (c) both combined. With this in mind, we process videos and transcriptions of the Public DGS Corpus [4] and focus on a subset with selected glosses as classes for the model. Since glosses are German words corresponding to the core meaning of a sign [5], we focus on cases where a single individual hand sign is labelled with multiple different glosses and the mouth actions are important for this disambiguation. The disambiguation of the ambiguous hand signs using the mouth actions is a goal of the recognition.

2 Related Work

In [6], the authors develop a framework to recognise mouthings in continuous SL using the RWTH-Phoenix corpus [7] and are, to the best of our knowledge, the first to apply dedicated viseme recognition for ASLR. Koller et al. also propose a method for automatic mouthing annotation for SL corpora [8] and a way to model mouth shapes [9] while outperforming state-of-the-art SL mouth shape classification for their time. However, mouthing, which refers to the production of words from the spoken language with the lips, represents only one type of mouth actions in DGS. There are also mouth gestures that do not correspond to spoken language words. [10] train a model to classify mouth gestures from videos of the Public DGS Corpus [4]. Their work on mouth gesture classification is extended while optimizing the preprocessing steps [11].

Several works on ASLR and ASLT explicitly employed the mouth area as an input feature and outperformed previous models on the RWTH-Phoenix corpus [2]. The STMC Network [12] is the current state of the art of ASLT for the RWTH-Phoenix corpus to the best of our knowledge. It employs the whole face as a feature, indicating that non-manuals may contribute to the state of the art. Since the face area contains the mouth, the model possibly learned from the mouth region as well. As far as we know, our paper is the first ASLR work for DGS that investigates the impact of modelling mouth actions in the context of ambiguous hand signs.

Automatic lip reading is a closely related task from the perspective of computer vision, as it also focuses on the mouth area. A model for word-level lipreading was trained in [13], achieving state-of-the-art results. In [14, 15], datasets for word-level lip reading for the German language were created while training models with them. We will use the architecture of [14] for our experiments too.

3 Method

3.1 Dataset

We processed the Public DGS Corpus [4], which provides SL videos (640x360px, 50 FPS) of fluent, including Deaf, signers from all around Germany with transcription and annotated glosses. We also used the online dictionary DW-DGS [16] that includes entries representing signs from the corpus. Each dictionary

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

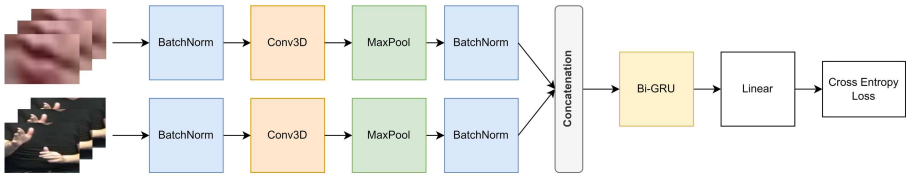

Fig. 1: Model for the last experiment

entry includes a concordance to corresponding gloss timestamps when the sign is used in the corpus. Using these resources, our researchers, including a fluent signer, went through the DW-DGS dictionary and selected entries where: (1) their concordance contains two glosses with different meanings and an amount of instances sufficient for training and (2) the manual signs of the two glosses are nearly identical in both hand form and movement, as manually determined by our team after utilising the corpus' original annotation to filter for these criteria. This allowed us to create 6 pairs of glosses from ambiguous signs (12 classes). For every pair, the two glosses have the same hand sign but different meanings and possibly different mouth actions. We selected this number of classes due to the high computational requirements, so that we can run experiments in a feasible amount of time. This also allowed us to manually inspect the instances and also ensure their validity. We prioritized the pairs of glosses that contained the biggest amount of instances, allowing more robust training and reliable evaluation.

To balance the gloss recognition for every gloss pair, we randomly removed instances of the gloss in a pair with more occurrences. This ensures that glosses from the same hand sign have equal number of instances. As a result, our data consists of 2948 instances for all 12 classes. The final frame of each video was duplicated and repeatedly concatenated to the end of the video, so that a standardized length of 28 frames for each video is achieved. We randomly split the data into three sets - training, validation, and test - in an 8:1:1 ratio while ensuring that the distribution of the classes was preserved in each of the sets. The split was performed without considering the signers, resulting in potential overlap of signers across all sets. There is a multitude of signers in every set, therefore we don't expect the model to learn features specific to individual signers.

3.2 Experiments

We implemented[1] a neural network consisting of Conv3D, bidirectional GRUs and linear layers (Fig. 1). Past works demonstrated that Conv3D and BiGRU layers can achieve state-of-the-art results for automatic lip reading [13, 17]. We use the architecture implemented in [14], since it showed promising performances in classifying mouth videos articulating German words. In the first experiment, the mouth region in each frame of the video clips is extracted, scaled to 150x100

[1]The code can be found at `https://github.com/NPhamDinh/AmbiguousDGS`. This includes a script to extract our dataset from the Public DGS corpus.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

ROI	Validation Accuracy	Test Accuracy
upper body (hands)	62.7%	63.3%
mouth	44.9%	40.7%
mouth + upper body (hands)	**69.9%**	**68.0%**

Table 1: The performance of the model for every region of interest on the validation and test set.

pixels, normalised and fed into the model. In the second experiment, the same is done with the region containing the upper body including the hands. These two inputs are used in the third and final experiment, where they are fused in the model following the Conv3D layer. In all layers, a dropout of 50% is being applied and the ReLu activation function is used, similar to [14, 17]. The Adam optimizer is used with the initial learning rate of 10^{-5}. In each experiment, we trained the model for 5,000 epochs with a batch size of 32 and used the weights with the highest validation accuracy for testing. In each epoch, RandAugment [18] is applied on the input. The first two experiments each took approximately 3 days to finish while the last experiment ran for 7 days. For the training, a NVIDIA GeForce GTX 1080 Ti was used with a memory of 24 GB for the first two experiments and 42 GB for the last one.

4 Results

Gloss (Translation)	No. of Instances	F1-score			Pairwise False Negatives		
		upper body (hands)	mouth	upper body (hands) + mouth	upper body (hands)	mouth	upper body (hands) + mouth
FERTIG1A (finished)	344	60.0%	36.4%	**66.7%**	4.3%	11.4%	**3.7%**
SCHON1A (already)	344	61.3%	45.2%	**74.4%**	3.3%	5.7%	**1.7%**
GEHÖREN1* (belong)	303	57.7%	15.2%	**58.6%**	**2.0%**	12.9%	**2.0%**
MEIN1 (my)	303	**81.2%**	49.2%	80.0%	**1.0%**	9.7%	1.7%
GUT1 (good)	85	12.5%	0.0%	**21.1%**	0.7%	11.1%	**0.0%**
SCHÖN3 (nice)	85	**53.3%**	0.0%	33.3%	**1.0%**	11.1%	1.7%
WAR1 (was)	277	**69.0%**	40.7%	61.5%	**1.3%**	7.1%	2.3%
FRÜHER1* (earlier)	277	65.4%	24.1%	**68.9%**	3.0%	**0.0%**	1.3%
NUR2A (only)	370	64.9%	63.2%	**68.9%**	4.0%	10.8%	**2.0%**
WENN1A (if)	370	65.0%	67.6%	**77.8%**	2.7%	18.9%	**2.0%**
GLEICH1A* (even)	95	47.6%	30.8%	**60.0%**	0.3%	**0.0%**	**0.0%**
WIE3A (like)	95	**66.7%**	11.1%	**66.7%**	1.0%	**0.0%**	1.0%

Table 2: Performance of the models for the glosses in the test set. The glosses with the same number of instances share the same hand sign. For each row, the highest F1-score and lowest pairwise false negative is bold. Pairwise false negative measures the percentage of test instances that were predicted as the other gloss sharing the same sign, instead of the actual class in the row.

The average accuracy over all classes is shown in Table 1. The model trained on both mouth and hands area achieved the highest accuracy, suggesting that

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

adding the mouth area as input can significantly improve models. Using only the mouth surprisingly achieved an accuracy of 40.7%, which further underlines how useful the information provided by the mouth area is to differentiate signs.

In Table 2 we present per-class results. Per-class F1-scores, computed over all classes, are equal or better by adding the mouth area, with the exception of 3 classes. When it concerns the percentage of false negatives computed for every pair of ambiguous hand signs, one can see that for 3 classes, full disambiguation was achieved by using only the mouth. For another 6 classes, combining upper body/hands area and mouth reduced pairwise false negatives over just using upper body/hands area.

In an effort to explain why the inclusion of the mouth area did not always contribute, we performed manual inspection of the video samples. It was found that often mouthings do not accompany the signs, since they may be disambiguated through the context. Furthermore, the DGS Corpus videos had a resolution of 640x360 pixels, resulting into compromised video quality for the extracted mouth area. Additionally, the small amount of training instances per gloss may have been another reason for the relatively low acccuracies and F1-scores.

5 Conclusion

In order to assess the importance of including the mouth area for ASLR systems, we trained a deep learning model with (a) the upper body/hands area, (b) the mouth and (c) both combined as input. The model combining hands and mouth as input achieved the best test accuracy and performed better in disambiguating hand signs for most of the cases. The results give insights into how useful the mouth region can be for ASLR and demonstrate that modelling the mouth area is important, especially for ambiguous hand signs.

Further work should optimize the hyper-parameters of the networks separately and consider the role of context. Then, we aim to focus on efficient ways of increasing the amount of classes. The modelling of the mouth area could be incorporated into the state-of-the-art ASLR and ASLT systems. Additionally, this research could also be extended by investigating ambiguous signs of SLs other than DGS and exploring the benefits of utilizing other non-manual features, such as eye gaze, blinks, cheeks, shoulders or head movements.

6 Acknowledgement

This research was funded by the German Ministry of Research and Education (BMBF) under the project SocialWear and the Agence Nationale de la Recherche (ANR) and the Deutsche Forschungsgemeinschaft (DFG) under the trilateral ANRDFG-JST call for project KEEPHA.

References

[1] Adil Er-Rady, R. Faizi, R. Oulad Haj Thami, and H. Housni. Automatic sign language recognition: A survey. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–7, May 2017.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

[2] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[3] Susanne Mohr. *Mouth Actions in Sign Languages*. De Gruyter Mouton, Boston, 2014.

[4] Reiner Konrad et al. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release, 2020.

[5] Reiner Konrad et al. Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation Conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany, 2022.

[6] Oscar Koller, Hermann Ney, and Richard Bowden. Read My Lips: Continuous Signer Independent Weakly Supervised Viseme Recognition. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 281–296, Cham, 2014. Springer International Publishing.

[7] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December 2015.

[8] Oscar Koller, Hermann Ney, and Richard Bowden. Weakly supervised automatic transcription of mouthings for gloss-based sign language corpora. In *LREC 2014 - Ninth International Conference on Language Resources and Evaluation*, 2014.

[9] Oscar Koller, Hermann Ney, and R. Bowden. Deep Learning of Mouth Shapes for Sign Language. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 477–483, 2015.

[10] Maren Brumm, Ronan Johnson, Thomas Hanke, Rolf-Rainer Grigat, and Rosalee Wolfe. Use of Avatar Technology for Automatic Mouth Gesture Recognition. Poster presented at 2nd SignNonmanuals Workshop (SignNonmanuals 2), Graz, Austria, May 2019.

[11] Maren Brumm and Rolf-Rainer Grigat. Optimised preprocessing for automatic mouth gesture classification. In *12th International Conference on Language Resources and Evaluation, LREC*, pages 27–32. ELRA, 2020.

[12] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016, 2020.

[13] Huijuan Wang, Gangqiang Pu, and Tingyu Chen. A Lip Reading Method Based on 3D Convolutional Vision Transformer. *IEEE Access*, 10:77205–77212, 2022.

[14] Dinh Nam Pham and Torsten Rahne. Entwicklung und Evaluation eines Deep-Learning-Algorithmus für die Worterkennung aus Lippenbewegungen für die deutsche Sprache [Development and Evaluation of a Deep Learning Algorithm for German Word Recognition from Lip Movements]. *HNO*, 70(6):456–465, 2022. Erratum in HNO, 70(6):466-467, 2022 [19].

[15] Gerald Schwiebert, Cornelius Weber, Leyuan Qu, Henrique Siqueira, and Stefan Wermter. A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6829–6836, Marseille, France, June 2022. ELRA.

[16] Jana Löffler, Ronja Molzer, and Charis Jolanthe Rasch. Gender-Fair Language in the Translation of Signed Utterances in the DGS-Korpus Project – Relevance and Challenges. Project Note AP11-2020-01, DGS-Korpus project, Hamburg University, Germany, 2020.

[17] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[18] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020.

[19] Dinh Nam Pham and Torsten Rahne. Erratum zu: Entwicklung und Evaluation eines Deep-Learning-Algorithmus für die Worterkennung aus Lippenbewegungen für die deutsche Sprache. *HNO*, 70(6):466–467, 2022.