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 Berlin)

Introduction

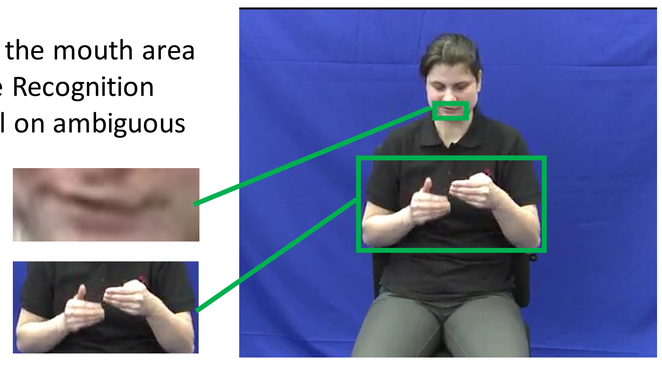
- Sign languages (SL) are multi-channeled languages, relying on visual-spatial components to communicate

Mouth actions

- Use of the mouth complements manual signing
- Limited research on the topic so far
- At least 3 functions:
  - Meaning specification
  - Sole carrier of meaning
  - Disambiguation (distinguishing homonyms like Schwester/Bruder; sister/brother in German SL)

Objective

- Evaluate the importance of the mouth area in Automatic Sign Language Recognition systems by training a model on ambiguous signs using:
  - (a) upper body and hands,
  - (b) mouth only,
  - (c) both combined.

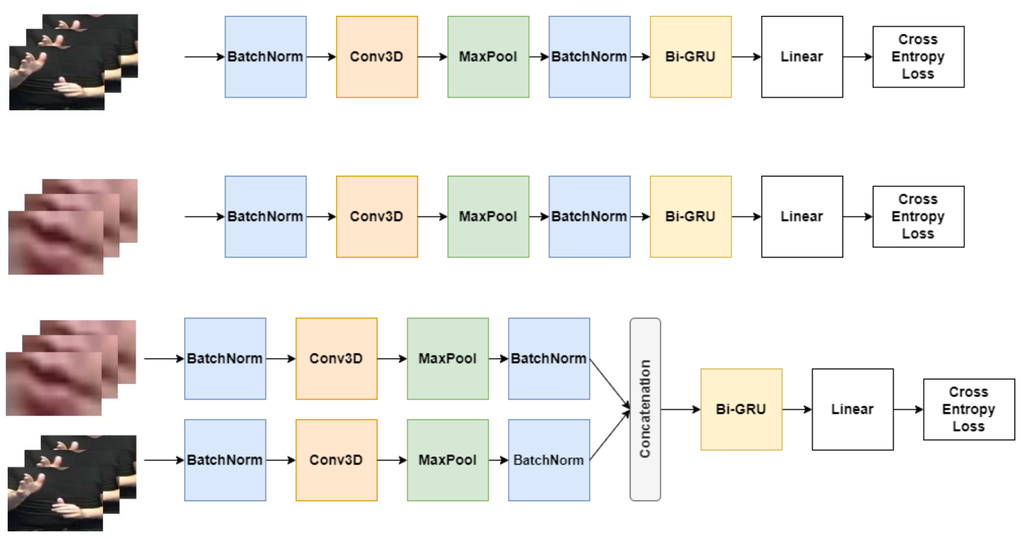


Preprocessing

- Mouth and upper body/hands extracted
- Scaled to 150x100px
- Fixed length of 28 frames by repeatedly appending the last frame
- Applied RandAugment on training set as data augmentation
- Pixel values normalised

Experiments

Model architecture



Experiment setup

- 5000 epochs for each experiment, best performing weights for validation and testing
- Batch size: 32, adam optimizer with initial learning rate of 10^{-5}
- NVIDIA GeForce GTX 1080 Ti
- Runtime of 3 days for first two experiments and 7 days for the last one

Discussion

- Model with combined input achieved highest accuracy, suggesting that adding the mouth area can improve models
- Mouth on its own surprisingly with decent accuracy of 40.7%, underlines usefulness of the mouth area to differentiate signs
- Inclusion of the mouth area did not always perform the best per class
- Possible reasons:
  - Mouthing didn't always accompany signs due to disambiguation by context
  - Low resolution (640x360px) of original videos results into poor video quality of the mouth area
- Small amount of instances per gloss could be reason for relatively low scores

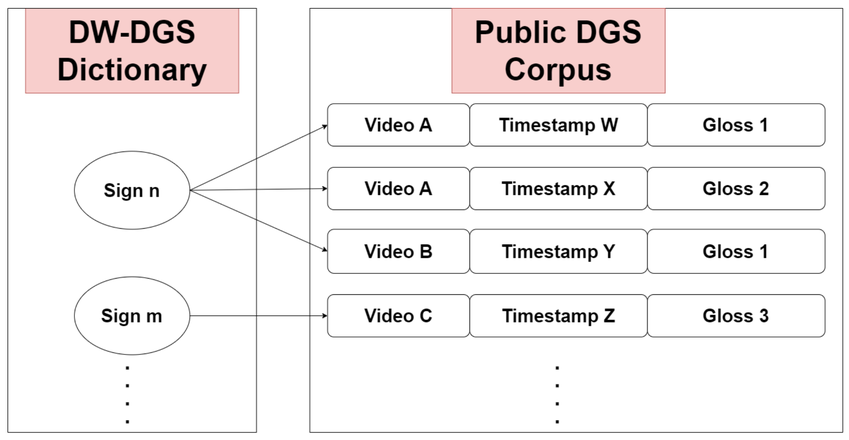
Supported by the German Federal Ministry of Education & Research (BMBF) through project SocialWear (grant 01IW20002) and Agence Nationale de la Recherche (ANR) and Deutsche Forschungsgemeinschaft (DFG) under the trilateral ANRDFG-JST call for project KEEPHA.

Dataset

Source

- Entries in the DW-DGS Dictionary[1] represent manual signs
- Includes concordance to DGS Corpus[2] with corresponding timestamps of glosses

→Occurrences of manual signs with different glosses (core meanings)



[1] dw-dgs.de
[2] https://doi.org/10.25592/dgs.corpus-3.0

Criteria for the selection of signs

- Concordance contains two glosses with:
  - Different meanings
  - Sufficient amount of instances for training
- Manual signs of the two glosses are indeed nearly identical in both hand form and movement (manually checked)

Key details

- 12 classes: 6 pairs of two glosses from the same sign
- Fluent, including Deaf, signers from all around Germany
- 2948 instances, 640x360px, 50 FPS
- Ensured equal instance numbers for gloss pairs by random removal from the gloss with more instances
- Training-validation-test split with 8:1:1 ratio

Results

Accuracies of the model for the regions of interests

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%**

Performance for the glosses in the test set

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%

Conclusion and outlook

- Model combining hands and mouth as input achieved best test accuracy and performed the best in disambiguating hand signs
- Results give insights into how useful the mouth region can be for ASLR
- Consider role of context for further work
- Possible incorporation of modelling of mouth area into state-of-the-art ASLR and ASLT systems
- Extend for sign languages other than DGS
- Explore benefits of utilizing other non-manual features, such as eye gaze, blinks, cheeks, shoulders or head movements

Contact details
Dinh Nam Pham, dinh_nam.pham@dfki.de
Vera Czehmann, vera.czehmann@dfki.de
Eleftherios Avramidis, eleftherios.avramidis@dfki.de