Signer Independent Viseme Recognition with Neural Networks

Arild Matsson
University of Gothenburg
(Main contributor)

Mehdi Ghanimifard
University of Gothenburg
(Contributions marked in [square brackets])

Abstract

Mouthing is an important part of sign languages, which partially serves to disambiguate between similar lexical signs. In automatic sign language recognition (ASLR), this calls for sequential classification of mouth shapes as visemes. We build upon previous work on this task, and explore the extended use of neural networks, using ResNet50 and LSTM. We develop three different models which partially surpass previous work. An even greater performance is expected from allowing more training time.

1 Introduction

Sign languages are different in modality from spoken languages. The shape, position and movement of hands is generally the most significant modality. Other modalities include mouthing, eyebrows and other facial features, as well as torso and head position. Mouthing is typically used either for grammatical markers, or lexically in conjunction with hand movement. For example, when signing the verb ”write”, a signer would typically mouth “O-A-I-T”, but substituting it with mouthing ”TH” would correspond to adding the adverb ”sloppily”. When used lexically, mouthing is redundant or distinguishing to varying extent, as sets of signs can share the same hand movement but differ in mouthing.

Fisher (1968) introduces “visemes” as sets of phonemes that are identical in visual appearance. In sign languages, the lexical mouthing of a word is generally derived from how the corresponding word is produced in spoken language. That is, the *viseme* sequence is copied from a word in spoken language. In practice, however, this sequence is often simplified and/or reduced. For instance, the Swedish word “medlem” (*member*) uses six visemes when spoken, while the corresponding sign in Swedish sign language (STS) is typically mouthed using only three (“m-e-m”).

Mouthing has been identified as an an essential and potentially distinguishing part of sign languages. Nevertheless, research on mouthing in the context of ASLR has not been extensively conducted (Antonakos et al., 2015).

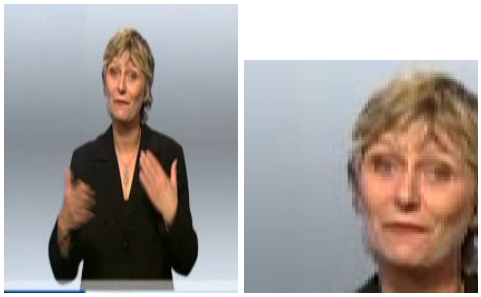
The present work aims to improve state-of-the-art sequential classification of mouth shapes as visemes. In section 1.1, we summarize previous work that this builds upon. The dataset and preprocessing is presented in section 2, and our own models are presented in section 3. Results are shown in section 4 and a concluding discussion in section 5.

1.1 Related Work

This work is largely based on Koller et al. (2014), which both introduces a viseme-annotated dataset and presents an ASLR approach based on it. They use facial landmarks for features and train a Hidden Markov Model (HMM) model.

A follow-up work uses deep learning for mouth shape classification, replacing facial feature extraction with a convolutional neural network (CNN). After classification, decoding the sequence is done using an HMM model, similar to the previous work. (Koller et al., 2015).

In the present work, we build upon this and explore the potentials of further extending the use of neural networks in the model. We do this using Long Short-Term Memory (LSTM) neural networks (Hochreiter and Schmidhuber, 1997) building on top of ResNet50 (He et al., 2015). [These were mainly Mehdi’s ideas, but discussed together.]



(a) Original (b) Cropped

Figure 1: Example frame from dataset.

2 Dataset

We use the dataset by Koller et al. (2014), which contains 3687 partly sequential video frames, annotated for visemes by five annotators. It is divided into 35 sentences, within each of which the frames are sequential. The frames are a part of the RWTH PHOENIX Weather dataset (Forster et al., 2012) recorded from TV-broadcast weather reports in German Sign Language (DGS). An example frame is shown in Figure 1a.

Note that a sentence in the context of this work is not a sequence of words, but of frames; word boundaries are ignored.

2.1 Preprocessing

Images were converted using the pre-trained ResNet50 (He et al., 2015). Only the first 49 layers of the model were used, thus every image was converted to a $7\times7\times2048$ tensor which was considered as features for input. [Mehdi showed how to modify the ResNet50 model like this.]

The annotations were converted to one-hot vector representations over the vocabulary of 40 phonemes. Some frames are annotated with more than one phoneme, in which case they are represented as the average of one-hot vectors. [Mehdi's idea.]

3 Models

We developed and tested three different models. All models are neural networks with a 32-unit LSTM (Hochreiter and Schmidhuber, 1997) layer. Here, the 3D tensor is put in its sequential context. Subsequently, a 40-unit dense layer with Softmax activation serves for classification. Each frame, in its sequential context, is thus mapped to one of 40 viseme labels.

Each model is a variation of this common form. They vary in neural network structure as well as preprocessing steps. The variations, named M1, M2 and M3, are detailed in the following subsections.

Implementations are available at `https://github.com/arildm/eslp-mouthing`, in commits tagged `m1`, `m2` and `m3`, respectively. [The `division-of-work` branch contains in-code comments indicating lines of code that were written by Mehdi.]

3.1 Model 1: M1

The model is essentially described above, with the difference that input 3D tensors and output vectors are converted to bigrams. Thus, a sample input is a sequence of two frame representations, and the output is the viseme class vector for the second one.

3.2 Model 2: M2

The dataset is now split by sentence. Each split is padded with zeroes to equal length (156, the length of the longest sentence). A sample no longer corresponds to a frame bigram, but to a 156-long sequence. [Mehdi's idea to go beyond bigrams.]

Input 3D tensors are cropped to focus more on the face part of frame images. The cropping reduces the 7×7 data to the top-center 3×3 part. [Mehdi wrote code for cropping.]

3.3 Model 3: M3

Aiming to reproduce the preprocessing done by Koller et al. (2014), this model includes augmentation of frame image input. First, images are cropped to an area containing the face, and resized to restore the aspect ratio which is distorted in the original dataset (Figure 1b). Augmentation is then done by randomly applying slight rotation and shifting to the cropped images. This preprocessing is done in a separate step before ResNet50 conversion. [Mehdi's idea, and he wrote the code for the data augmentation.]

The data augmentation results in 10 new training sequences for every original sentence. In addition to larger training data, this allows for a separate test set which contains the straight and unshifted images (but still cropped and resized).

4 Results

Models were trained until convergence. They were then evaluated using the evaluation script included in the RWTH PHOENIX Weather dataset package. It measures precision and recall against the

Model	Precision	Recall	F-score
Koller et al. (2014)	31.3	**43.2**	32.7
M1	38.7	2.5	4.7
M2	**57.1**	32.2	**41.2**
M3	43.5	10.4	16.8

Table 1: Precision and recall in [%].

A	E	F	I	L	O	P	Q	S	T	U	GB
31.8	-	-	-	-	7.3	-	-	-	-	-	8.0
10.0	14.3	-	-	-	-	-	-	-	-	-	8.0
3.8	-	-	-	-	-	-	-	-	-	-	8.9
4.9	-	-	-	-	8.9	-	-	-	-	-	9.5
6.4	-	-	-	-	-	-	-	-	-	-	3.6
6.8	-	-	-	-	45.5	-	-	-	-	-	13.5
4.6	19.0	-	-	-	-	-	-	-	-	-	10.2
9.6	-	-	-	-	2.4	-	-	-	-	-	9.7
2.0	-	-	-	-	2.4	-	-	-	-	-	4.4
15.7	66.7	-	-	-	30.1	-	-	-	-	-	17.0
4.4	-	-	-	-	3.3	-	-	-	-	-	7.2
-	-	-	-	-	-	-	-	-	-	-	-

Table 2: Confusion matrix in [%] for M3 with true classes on the y-axis and predicted classes on the x-axis.

annotation file which is also included. The results of this evaluation are shown in Table 1. They are compared to the baseline by Koller et al. (2014).

M2 reaches a significantly higher precision that the baseline (57.1%). None of the present model surpasses baseline recall, although M2, again, comes closest (32.2%) and also reaches the highest F-score at 41.2%.

The low recall of the present models indicates conservativity. This is indeed confirmed by the confusion matrix in Table 2, where two classes are clearly preferred and only four are ever selected at all.

5 Discussion

Approximately 80% of the rather expensive ResNet50 conversion output was thrown away when cropping in M2. This was remedied in M3 by moving cropping to before conversion. But in fact, the reduced input size in M2 has a drastic influence on the model size (2.3M trainable parameters as compared to 12.9M). This difference reduces the risk of underfitting, and is likely to explain the poor performance of M1 and M3.

Between M1 and M3, a significant improvement was provided by the latter. M3 features data augmentation as well as longer sequences, and it is difficult to tell which of these were more important for the improvement.

It is notable that M3 features a train/test split that is not present in the two prior models, making them likely to perform worse on new data.

Some final, brief suggestions for improvements are given below:

Segmentation Replacing our imprecise fixed-coordinates cropping with intelligent detection and segmentation of the mouth area.

Inception Replacing ResNet50 with Inception for closer correspondence to Koller et al. (2015).

CNN fine-tuning Connecting ResNet50 (or Inception) directly to the trainable model, fine-tuning it to the data at hand.

References

E. Antonakos, A. Roussos, and S. Zafeiriou. 2015. A survey on mouth modeling and analysis for sign language recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Lubljana, Slovenia, pages 1–7.

Cletus G. Fisher. 1968. Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research* 11(4):796–804. https://doi.org/10.1044/jshr.1104.796.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. 2012. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Language Resources and Evaluation*. Istanbul, Turkey, pages 3785–3789. http://www.lrec-conf.org/proceedings/lrec2012/pdf/844_Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR* abs/1512.03385. http://arxiv.org/abs/1512.03385.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Oscar Koller, Hermann Ney, and Richard Bowden. 2014. Read my lips: Continuous signer independent weakly supervised viseme recognition. In *Proceedings of the 13th European Conference on Computer Vision*. Zurich, Switzerland, pages 281–296.

Oscar Koller, Hermann Ney, and Richard Bowden. 2015. Deep learning of mouth shapes for sign language. In *Third Workshop on Assistive Computer Vision and Robotics, ICCV*. Santiago, Chile, pages 477–483.