Neural Models for Fingerspelling Recognition from Video

Bowen Shi, Taehwan Kim, Jonathan Keane, Weiran Wang, Hao Tang,
Gregory Shakhnarovich, Diane Brentari, Karen Livescu
{bshi,weiranwang,haotang,greg,klivescu}@ttic.edu,
taehwan@caltech.edu,{dbrentari,jonkeane}@uchicago.edu

Introduction While there has been extensive research on automatic speech recognition, much less progress has been made for sign languages. We study the problem of recognizing sequences of fingerspelled letters in videos of American Sign Language (ASL). Fingerspelling is often used for names and borrowed words from English or other languages (see examples in Fig. 1). Recognizing it is challenging because it involves quick, coarticulated motions and exhibits significant inter-signer variation. We have collected a data set of fingerspelling videos and have developed several types of recognition models based on deep neural networks.

Most previous related work has focused on restricted conditions such as isolated letters, small vocabularies, or signer-specific modeling (Liwicki and Everingham 2009; Ricco and Tomasi 2009). In such settings, letter error rates of 10% or less have been obtained. In contrast, we focus on *lexicon-free* recognition of fingerspelling produced by *multiple signers*. This abstract summarizes our work to date.[1]

Methods Because of the speed of motion in fingerspelling, very few frames look like the canonical letter handshapes. We therefore develop sequence models that account for the dynamics of fingerspelling. We consider several models inspired by successful models for speech recognition, but customized for the fingerspelling task in various ways:

- A traditional hidden Markov model (HMM)-based approach, taking as observations neural network classifier predictions of letters and handshape features.
- A segmental conditional random field (SCRF), with feature functions based on summarizing neural classifier predictions over hypothesized letter segments. The SCRF is computationally demanding, so we also consider using it to rescore the top hypotheses of the HMM-based model.
- A neural attention model. This is an encoder-decoder recurrent neural network-based model, where at each time step the decoder has the ability to access arbitrary portions of the input (through an attention mechanism) in determining the next letter label.

Results We study several settings: signer-dependent (models trained and tested on the same signer), signer-independent (models trained on several signers and tested on another), and signer-adapted (warm-starting with signer-independent models and fine-tuning using a small amount of test signer data). Tab. 1 shows the results on test signers. In the signer-dependent case we obtain comparable error rates to previous approaches that used a constrained lexicon, although we recognize *unconstrained* letter sequences. In the signer-dependent case the SCRF performs best, possibly because the many segment-level feature functions can be well-tuned to the signing habits of the test signer. In the signer-independent and adapted cases, however, the neural attention model does best; this model is in some sense the simplest and is trained end to end, so it likely benefits from larger amounts of training data.

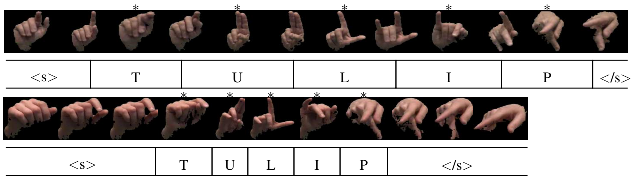

Figure 1: Subsampled frames, after hand segmentation, from the fingerspelled word TULIP produced by two signers. The most canonical frame for each letter is marked (*).

We are currently collecting a much larger data set of fingerspelling "in the wild"—from online social and news media—so as to extend our work to a greater variety of signers and visual conditions, as well as to learn to jointly detect and transcribe fingerspelling within running ASL.

	HMM	Rescoring SCRF	1^{st}-pass SCRF	Attn.
Signer-ind.	57.2	55.3	60.6	50.3
Adapted	33.6	32.0	30.3	29.1
Signer-dep.	14.6	11.5	8.8	12.5

Table 1: Mean letter error rates (%) over four test signers.

References

Kim, T. *et al.*. 2016. Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation. *arXiv preprint arXiv:1609.07876*.

Liwicki, S., and Everingham, M. 2009. Automatic recognition of fingerspelled words in British Sign Language. In *2nd IEEE Workshop on CVPR for Human Communicative Behavior Analysis*.

Ricco, S., and Tomasi, C. 2009. Fingerspelling recognition through classification of letter-to-letter transitions. In *ACCV*.

[1] Portions of this work have appeared in (Kim 2016).