Fingerspelling recognition in the wild with iterative visual attention

Bowen Shi[1], Aurora Martinez Del Rio[2], Jonathan Keane[2], Diane Brentari[2], Greg Shakhnarovich[1], and Karen Livescu[1]

[1] Toyota Technological Institute at Chicago, USA
{bshi,greg,klivescu}@ttic.edu
[2] University of Chicago, USA
{amartinezdelrio,jonkeane,dbrentari}@uchicago.edu

Abstract. We address the problem of fingerspelling sequence recognition in American Sign Language (ASL) videos collected in the wild, mainly from YouTube and Deaf social media. Most previous related work has focused on controlled settings (studio environment, limited number of signers). Our work aims to address the challenges of real-life data, where hand detection and segmentation is challenging. We propose an end-to-end model based on an iterative attention mechanism, without explicit hand detection or segmentation, that dynamically focuses on increasingly high-resolution regions of interest. It outperforms prior work by a large margin. We also introduce a newly collected data set of crowdsourced annotations of fingerspelling in the wild, and show that performance can be further improved by training on this additional data set.

Keywords: American Sign Language, Fingerspelling, Attention

1 Introduction

We study the problem of American Sign Language (ASL) fingerspelling recognition from naturally occurring online sign language videos. Fingerspelling is a component of ASL in which words are signed letter by letter, using an alphabet of canonical letter handshapes, and is used frequently for names and other content words. Most prior related work has focused on data collected in a controlled environment [1, 3]. Compared to studio data, naturally occurring fingerspelling images often involve more complex visual context and more motion blur, especially in the signing hand regions (see Figure 1). Thus hand detection, an essential pre-processing step in typical recognition pipelines [3, 4], becomes more challenging.

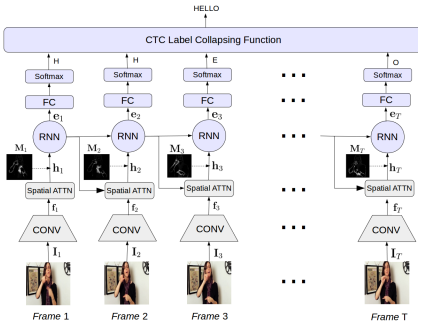
We propose an approach for fingerspelling recognition that does not rely on hand detection, based on an attention-based model trained end-to-end from raw image frames. In addition, we introduce a new, publicly available data

Fig. 1: Fingerspelling in studio data (leftmost image) vs. in the wild.

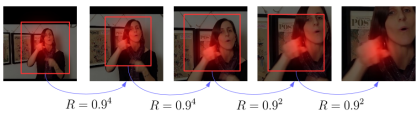
set (`https://ttic.edu/livescu/ChicagoFSWild.htm`) of crowdsourced fingerspelling video annotations, and show that training on it leads to significantly improved fingerspelling recognition. This abstract is based on the prior work [5].

2 Task and model

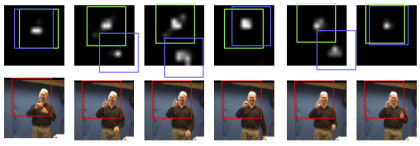
The fingerspelling recognition task takes as input a sequence of image frames (or patches) $\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_T$ and produces as output a sequence of letters $w = w_1, w_2, ..., w_K$, $K \leq T$.



(a) Recurrent CNN with attention



(b) Iterative attention; R=zoom factor



(c) One iteration of iterative attention

Fig. 2: An overview of the model.

Our model is based on a convolutional recurrent architecture (see Figure 2a). At time step t, a fully convolutional neural network is applied on the image frame $\mathbf{I}_t$ to extract a feature map $\mathbf{f}_t$, which is transformed into a visual feature vector $\mathbf{f}_{tij}$: $\mathbf{h}_t = \sum_{i,j} \mathbf{f}_{tij} A_{tij}$ through an attention mechanism:

$$v_{tij} = \mathbf{u}_f^T \tanh\left(\mathbf{W}_d \mathbf{e}_{t-1} + \mathbf{W}_f \mathbf{f}_{tij}\right), \beta_{tij} = \frac{\exp\left(v_{tij}\right)}{\sum_{i,j} \exp\left(v_{tij}\right)}, \mathbf{A}_t = \frac{\boldsymbol{\beta}_t \odot \mathbf{M}_t^\alpha}{\sum_{p,q} \beta_{tpq} M_{tpq}^\alpha} \tag{1}$$

The label sequence $w = w_1, w_2, ..., w_K$ is produced from the frame-level sequence via the connectionist temporal classification (CTC) "label collapsing function", and the model is trained with the CTC loss [2].

The attention mechanism enables the model to focus on informative regions, but high resolution is needed in order to retain sufficient information in the attended region. However, using large images can lead to prohibitively large memory footprints. We propose to iteratively focus on regions within the input image frames, by refining the attention map. Specifically, given a trained attention model $\mathcal{H}$ we use the sequence of attention maps $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_T$ to obtain a new sequence of images $\mathbf{I}'_1, \mathbf{I}'_2, \ldots, \mathbf{I}'_T$ consisting of smaller bounding boxes within the original images. Then a new model $\mathcal{H}'$ that takes $\mathbf{I}'_1, \mathbf{I}'_2, \ldots, \mathbf{I}'_T$ as input is trained. We iterate this process for S steps until ROI images of sufficiently high resolution are obtained. This iterative process generates S models.

The iterative attention process is illustrated in Figures 2b, 2c. In each iteration s, we assign a score a_t^i equal to the attention value to the box of size $R_s|\mathbf{I}^s|$

centered at each of the top k peaks in the attention map $\mathbf{A}_t$ of frame $\mathbf{I}_t^s$. Finding the sequence of bounding boxes consists of solving the optimization problem

$$\arg\max_{i_1,\ldots,i_T} \frac{1}{T}\sum_{t=1}^{T-1} sc(b_t^{i_t}, b_{t+1}^{i_{t+1}}) \tag{2}$$

where $sc(b_t^i, b_{t+1}^j) = a_t^i + a_{t+1}^j + \lambda * IoU(b_t^i, b_{t+1}^j)$ is the linking score between box b_t^i and b_{t+1}^j.

3 Experiments

We use two data sets: Chicago Fingerspelling in the Wild (ChicagoFSWild) [6], which was carefully annotated by experts; and a crowdsourced data set we introduce here, ChicagoFSWild+. Both contain clips of fingerspelling sequences excised from sign language video "in the wild", collected from online sources such as YouTube and `deafvideo.tv`. ChicagoFSWild+ includes 50,402 training sequences from 216 signers, 3115 development sequences from 22 signers, and 1715 test sequences from 22 signers, with no overlap in signers in the three sets. Compared to ChicagoFSWild, the crowdsourcing setup allows us to collect dramatically more training data with significantly less expert/researcher effort.

We consider the following scenarios for initial processing of the input frames: **Whole frame**, which uses the full video frame with no cropping; **Face ROI**, which uses a region centered on a face detection box, but 3 times larger; **Hand ROI**, which uses a region centered on the box resulting from a signing hand detector. As Table 1 shows (in green), our detector-free approach ("Ours + whole frame") improves over previous work that uses a hand detector [6], but also benefits from hand/face detectors if they are available. All tested models benefit significantly from the new data: The crowdsourced annotations in ChicagoFSWild+ may be noisier, but they are much more plentiful. In addition, the crowdsourced training data includes two annotations of each sequence, which can be seen as a form of natural data augmentation.

Method	**ChicagoFSWild**	**ChicagoFSWild+**
Hand ROI [6]	41.9	41.2
+new data	57.5	58.3
Ours+whole frame	42.4	43.8
+new data	57.6	61.0
Ours+hand ROI	42.3	45.9
+new data	60.2	61.1
Ours+face ROI	**45.1**	**46.7**
+new data	**61.2**	**62.3**

Table 1: Results on *ChicagoFSWild/test* and *ChicagoFSWild+/test*. Black: trained on *ChicagoFSWild* only. Green: trained on *ChicagoFSWild and ChicagoFSWild+*.

The iterative zooming gives a large performance boost over the basic model. Figure 3 shows how the accuracy and the input image evolve in successive zooming iterations. Though no supervision regarding the hand is used for training, the

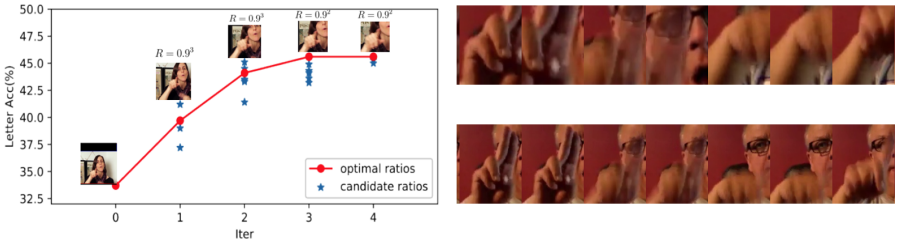


Fig. 3: Left: Letter accuracy vs. iteration in the Ours+face ROI setting. Right: Signing hands detected by the iterative attention detector (bottom row) vs. the off-the-shelf signing hand detector.

location of the signing hand is implicitly learned through the attention mechanism. Qualitatively, our model produces smoother hand tubes compared to the off-the-shelf hand detector used in [6] (see Figure 3).

4 Conclusion

We have developed a new model for ASL fingerspelling recognition in the wild, using an iterative attention mechanism, which does not rely on hand detection, segmentation, or pose estimation (but can benefit from hand/face detectors when available). Our model gradually reduces its area of attention while simultaneously increasing the resolution of its ROI, yielding a sequence of models of increasing accuracy. We also contribute a new data set of fingerspelling in the wild with crowdsourced annotations, which is larger and more diverse than any previously existing data set, and show that training on the new data significantly improves the accuracy of all models tested.

References

1. Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., Ney, H.: RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus. Language Resources and Evaluation pp. 3785–3789 (2012)
2. Graves, A., Fernandez, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: ICML (2006)
3. Kim, T., Keane, J., Wang, W., Tang, H., Riggle, J., Shakhnarovich, G., Brentari, D., Livescu, K.: Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation. Computer Speech and Language pp. 209–232 (2017)
4. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled. In: CVPR (2016)
5. Shi, B., Del Rio, A.M., Keane, J., Brentari, D., Shakhnarovich, G., Livescu, K.: Fingerspelling recognition in the wild with iterative visual attention. In: ICCV (2019)
6. Shi, B., Del Rio, A.M., Keane, J., Michaux, J., Brentari, D., Shakhnarovich, G., Livescu, K.: American Sign Language fingerspelling recognition in the wild. In: SLT (2018)