Project#8: Sign Segmentation in Indian Sign Language

Arqam Patel[1], **Arnav Singla**[2], **Mohit Kulkarni**[3], **Kumar Kanishk Singh**[4], **K Nikita**[5]
[1]210194, [2]210188, [3]190507 [4]210544 [5]180355
[1,2,4]SDS, [3]MTH, [5]CSE
{arqamrp21, arnavs21, mohitmk, kksingh21, nikitak}@iitk.ac.in

Abstract

Indian Sign Language has very little annotated or processed data, despite being one of the most commonly used sign languages in the world. Our main task, sign segmentation, refers to delineating temporal boundaries between sign language signs to distinguish individual signs in continuous signing footage. It is a preliminary step for building pipelines for automated sign language processing. These can be used to eventually build machine translation and generation systems for sign language, significantly boosting the accessibility of resources for the deaf community. We propose using transfer learning methods to achieve this in the absence of annotated data for the task.

1 Introduction

In natural language processing, we need to break down continuous streams of language into smaller units. In word segmentation, we identify word boundaries in a given sample. In many written languages, spaces are used to separate words. However, in other scripts, like Chinese and Japanese, that is not the case, which makes the task more challenging. Sign languages do not have separate symbols (signs) for denoting the end of lexical units. These rely on temporal and auxiliary cues like pause and articulation, as well as grammatical information, to convey this information.

Sign languages involve multiple modalities, including hand gestures, facial expressions and mouthing. This multimodal nature, as well as the lack of clear boundaries between signs, and variations in signing speed and style, make segmentation of continuous sign language difficult.

Due to the lack of annotated ISL data for sign segmentation, we propose to use a baseline model making use of inflational three-dimensional convolutional networks and temporal convolutional networks, trained on British Sign Language [8]. We use a semi-supervised transfer learning method called pseudo-labelling [7] that involves assigning 'pseudo-labels' to unlabelled data according to the baseline model. We assign pseudo-labels under the paradigm of source-free domain adaptation based on the assumption that sign segmentation in BSL is broadly similar to that in ISL. Along with this, we use changepoint detection, a technique that helps detect abrupt changes in gestural direction. We use a composition of the two categories to generate changepoint-modulated pseudolabels. We compare the efficacy of using pseudolabels, changepoints and CMPL labels as training data for the model.

2 Problem Definition

[8] A series of N video frames comprise the input, with $x = [x_1, ..., x_L]$.For each frame, we want to train the model to predict a binary output $y_i \in \{0, 1\}$. If the frame is inside a sign token, it is indicated by 0, otherwise, it is shown by 1. Then $y = [y_1, ..., y_N]$.[3]

Using a cross entropy loss and a truncated mean square loss, the model attempts to optimise the function, where the loss

$$L = \Sigma_{i=1}^{n} L_i$$

$$L_i = L_{CE} + \lambda L_{T-MSE}$$

where L_i represents the loss of each x, L_{CE} represents the cross entropy loss, L_{T-MSE} represents the truncated mean squared loss, and λ is a hyper-parameter used to specify the weight of the truncated mean squared loss.

3 Related Work

Considerable work has been done in word segmentation in text and speech. N-gram methods, Hidden Markov Model (HMMs) and Conditional Random Fields(CRFs) have previously been used with much success in cases of text segmentation. For languages like Chinese and Japanese, where space isn't used for word separation, CRFs segment the

words quite effectively. See [1] for an overview of the same.

Renz et al. [3] tackled the problem of sign segmentation using Temporal Convolutional Networks (TCN) to locate the temporal boundaries in sign language videos. In our task, long term-dependencies do not play much of a role; hence TCNs are suitable. Each Single-Stage TCN (SS-TCN) module is sequentially connected to each other and successively gives out better segmentation. One of the main contributions of this paper is their experiments on the generalization of this model trained on British Sign Language (BSL) to other sign languages. The authors found that their model was able to capture a lot of the shared variance of the two sign languages exploiting common features.

Building upon this work, the author in [4] proposed the use of semi-supervised learning techniques for better performance in languages where annotated data is scarce. The authors use the idea of source-free domain adaptation to fine-tune the model trained on BSL to German Sign Language (GSL). The model comprises two different blocks, one which produces the pseudo labels and changepoints, and the other combines them to fine-tune the original model.

Another work [5] looks at sign segmentation, taking into account the availability of subtitles with poorly segmented points. It uses a BERT-like encoder to encode text information from the subtitles and combines it with a MSTCN+I3D module. It is then passed through a transformer encoder that aligns the subtitle frames and sign-language video frames to produce segmentations.

4 Data Description

Despite the wide usage of Indian Sign Language, there is little to no video data available where the signs are annotated. Even if annotations are available, they are not aligned well with their respective signs. Following are the specifics of our dataset.

1. Training Set: We scraped unlabeled data from Deaf Enabled Foundation channel on Youtube: https://www.youtube.com/@deafenabledfoundation1117/featured. We got 722 videos of about 1 minute.

2. Test Set: We used ISL CSLTR dataset to get our test set. The CSLTR dataset has videos for 100 sentences in ISL performed by 7 different signers. We took two videos of each sentence and manually annotated it using VIA Sign Language Annotation Tool: https://github.com/RenzKa/VIA_sign-language-annotation[2]. We got 176 annotated videos.

5 Proposed Approach

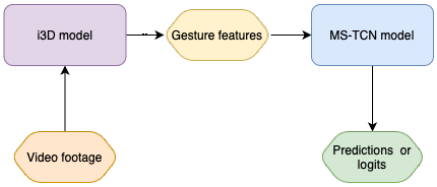

Figure 1: Model Architecture

After doing a thorough literature analysis to learn about previous efforts in Sign Language segmentation, we determined the approach for our work. We used the model described in [8] .The model uses an inflational 3D convolutional network to first extract spatio-temporal information from the frames, and then a multi stage temporal convolutional network to detect sign boundaries.

In order to boost accuracy, the authors of [7] use a technique termed source-free domain adaptation along with pseudo labelling and changepoint labelling. While this model was trained on BSL (British Sign Language), it was able to generalise to the Phoenix14 dataset on GSL.

Preliminary results suggested that the CMPL approach suffers from a high rate of false positives, and alleviating this would be a key task in order to improve the efficacy of the model.We first generate CMPL labels and using these label as ground truth we fine tune our model. To generate the CMPL labels, we convert the videos to tensor and pass them to i3D Model pre-trained on BSLCP1K corpus to get the features. These features were passed to the MSTCN Model to get Pseudo labels. We applied Changepoint Modulation Algorithm on these features to get the Changepoint labels. We combine these to get the CMPL labels.

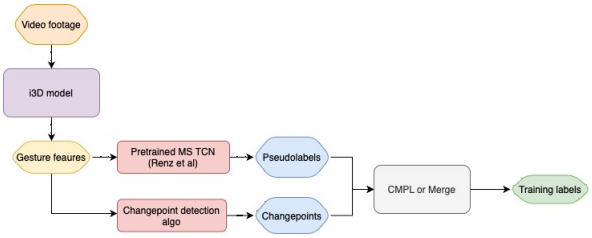

Figure 2: Training Strategy

We use the various labels (CP, PL, and CMPL) as training data to finetune the Model. To do so, we pass the videos to the i3D Model and get the features. We pass the features to MSTCN Model pre-trained on the BSLCP1K Corpus to get the predictions. We calculate the Cross Entropy Loss using these predictions and labels and update the parameters of the model using the Adam batch optimisation algorithm.

6 Experiments and Results

We compare the following strategies for constructing training sets to train the MS TCN model:

1. Changepoints: We solely use the feature extraction network to calculate the changepoints on the extracted features for the changepoint baseline.

2. CMPL: This strategy comprises two pseudo label transformations. The first transformation inserts new boundaries suggested by abrupt changes in feature space, when far from pseudo-label boundaries. The second refinement transformation aims to minimise potential bias towards the annotation style used.

3. Merge: This strategy takes the union of predictions and keeps all boundaries which are present either in the pseudo-labels or in the changepoint.

4. Pseudo-labels: We train the MSTCN model on our data to generate pseudo-labels.

Here are the training graphs for these strategies. Note that the graph labelled as test is for the validation set.

Figure 3: Training Graph for PL Model

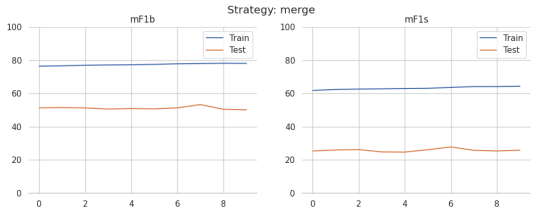

Figure 4: Training Graph for Merge Model

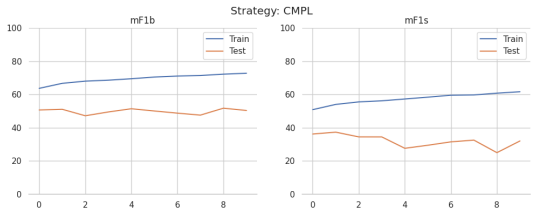

Figure 5: Training Graph for CMPL Model

We have provided the training graphs for these strategies in Figure 2, 3 and 4. Note that the graph labelled as test is for the validation set.

Here are some results on our test set comparing models trained using various kinds of train set construction strategies:

Strategy	**mF1B**	**mF1S**
Merge	20.38	7.59
CMPL	33.38	7.81
Pseudo labels	19.5	9.13

We can see that, as noted in the paper [7] CMPL is the best training set construction strategy for finetuning MS TCN models on a new sign language.

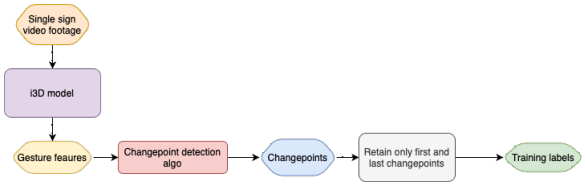

Figure 6: Proposed novel strategy

We also experimented with a novel data augmentation strategy involving use of lexicon-based datasets, in order to build recall for specific signs, as well as counter the tendency towards oversegmentation. We used the initial and final changepoints in each video (consisting of a single sign) as the true sign boundaries. Using a dataset of 640 single sign videos, we further finetuned the

model for multiple epochs. This slightly improved the mF1S score while significantly impacting the mF1B score.

Here is the link for our GitHub Repository: https://github.com/arqamrp/sign-segmentation-isl

7 Error Analysis

We evaluated our model using the same metric as [8]. The authors of this paper define two metrics **mF1B** and **mF1S** for evaluating boundary prediction and Sign Prediction respectively.

mF1B: A boundary prediction is defined to be correct, if its distance from the ground truth boundary is found to be less than a given threshold. We calculate the F1 scores for the boundaries. We use all integer-valued thresholds that fall within the closed interval [1, 4] and report the mean across thresholds, which is referred to as mF1B.

mF1S: The sign segments with an intersection over union higher than a given threshold are defined as correct. We average the F1 scores for thresholds from 0.4 to 0.75 with a step size of 0.05 and report the result as mF1S.

As we can see from the table of results, our model currently struggles with a high rate of false positives (i.e. oversegmentation).

8 Future Directions

There was a scarcity of annotations in data on Indian Sign Language. There is a need to build a properly annotated, diversely sourced dataset on the word level for both training and testing. Proper annotation of the dataset will also help increase the accuracy of sign language recognition systems.
Instead of the top down approach to segmentation that we tried, a bottom-up approach can also be explored. Something analogous to a sentence piece tokeniser can be built that pieces together different purely changepoint-separated segments into one coherent piece that represents one meaningful sign. This could potentially overcome the limitations of the current models by focusing more on linguistic features that spatiotemporal ones.

9 Individual Contribution

The contributions of each member are as follows:

Arnav Singla	Literature review of [8] and data scraping, Training, Evaluation, Data cleaning
Kumar Kanishk Singh	Literature review [5] and Data annotation, report
Arqam Patel	Literature review [7], Datasets survey [4] [6], Experimental training strategy, evaluation
K Nikita	Literature review of [3] , Data annotation, Evaluation
Mohit Kulkarni	Literature review [7] , Training, Evaluation, Data cleaning [1]

10 Conclusion

The results of running the MSTCN + i3D Model on Indian Sign Language were promising, although the model did struggle with oversegmentation. The best approach to generating pseudo training data was found to be CMPL, corroborating the findings of the paper by Ren et al [7]. Further training using the current approach, and more data

References

[1] Hannah Bull et al. *Aligning Subtitles in Sign Language Videos*. 2021. DOI: 10.48550/ARXIV.2105.02877. URL: https://arxiv.org/abs/2105.02877.

[2] Abhishek Dutta and Andrew Zisserman. "The VIA Annotation Software for Images, Audio and Video". In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19. Nice, France: ACM, 2019. ISBN: 978-1-4503-6889-6/19/10. DOI: 10.1145/3343031.3350535. URL: https://doi.org/10.1145/3343031.3350535.

[3] Yazan Abu Farha and Juergen Gall. "MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation". In: 2019.

[4] Abhinav Joshi et al. "CISLR: Corpus for Indian Sign Language Recognition". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10357–10366. URL: https://

aclanthology.org/2022.emnlp-main.707.

[5] Medet Mukushev et al. “Towards Semi-automatic Sign Language Annotation Tool: SLAN-tool”. In: *SIGNLANG*. 2022.

[6] Elakkiya R and NATARAJAN B. *ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition*. 2021. DOI: 10.17632/kcmpdxky7p.1.

[7] Katrin Renz et al. “Sign Segmentation with Changepoint-Modulated Pseudo-Labelling”. In: *Workshop on ChaLearn Looking at People Sign Language Recognition in the Wild, CVPR*. IEEE, 2021.

[8] Katrin Renz et al. “Sign Segmentation with Temporal Convolutional Networks”. In: *International Conference on Acoustics, Speech, and Signal Processing*. 2021.