Automatic Segmentation of Sign Language into Subtitle-Units

Hannah Bull, Michèle Gouiffès, Annelies Braffort

Automatic Segmentation of Sign Language into Subtitle-Units

Hannah Bull[1,2][0000−0002−9649−357X], Michèle Gouiffès[1,2][0000−0002−7152−4640], and Annelies Braffort[1][0000−0003−4595−7714]

[1] LIMSI-CNRS, Campus universitaire 507, Rue du Belvedère, 91405 Orsay, France
{hannah.bull, michele.gouiffes, annelies.braffort}@limsi.fr
[2] University of Paris-Saclay, Route de l'Orme aux Merisiers - RD 128, 91190 Saint-Aubin, France

Abstract. We present baseline results for a new task of automatic segmentation of Sign Language video into sentence-like units. We use a corpus of natural Sign Language video with accurately aligned subtitles to train a spatio-temporal graph convolutional network with a BiLSTM on 2D skeleton data to automatically detect the temporal boundaries of subtitles. In doing so, we segment Sign Language video into subtitle-units that can be translated into phrases in a written language. We achieve a ROC-AUC statistic of 0.87 at the frame level and 92% label accuracy within a time margin of 0.6s of the true labels.

Keywords: Sign Language, Segmentation, Sentence, Subtitle, Graph Neural Network, Skeleton Keypoints

1 Introduction

Sign Language (SL) is an essential means of communication for Deaf communities. SLs are visuo-gestual languages with no written form, instead using hands, body pose and facial expression as the medium of transmission. A natural way of recording SL is through video. The uniqueness of transmission medium, structure and grammar of SL requires distinct methodologies.

The treatment of language as a sequence of words from a lexicon is unsuitable for SLs [10]. The notion of a 'word' in SL is ill-defined, as the beginning or end of a sign in fluent discourse is unclear. Moreover, signs can occur simultaneously, further blurring the notion of a 'word' and rendering impossible the modelisation of SL as a linear sequence of words. The iconicity of SLs means that signs are strongly modified according to context and meaning, rather than being drawn largely unmodified from a lexicon.

Classic natural language processing tasks including speech-to-text, word embeddings and parts-of-speech tagging currently do not have direct counterparts in SL processing. Tasks such as automatic translation between SL and written language are in a preliminary stage, with translation only possible for short and rudimentary phrases with limited vocabulary [3].

We wish to define a sentence-like unit that can be used to segment SL into short and coherent sequences that can be translated individually. This task of segmentation of SL video is useful for numerous tasks, including software for subtitling assistance, reducing sequence length for continuous SL recognition, or phrase-level alignment between SLs and spoken or written languages. Manual segmentation of SL video into sentence-like units is a fastidious and extremely time consuming task, and so we aim to automatise this problem.

We define a *subtitle-unit* (SU) as a segment of SL video corresponding to the temporal boundaries of a written subtitle in accurately subtitled SL video. The SU is of linguistic relevance, as the person subtitling the SL video purposefully aligns phrases of text with what they consider to be equivalent phrases in SL. Implicitly, the subtitiler labels segments of SL video that can be translated into a phrase in written language.

Our key contribution is to present baseline results of the new task of automatically segmenting SL video at a sentence-like level. Our method is an adaptation of a state-of-the-art graph-based convolutional network for sequences of 2D skeleton data of natural SL. We also study the influence of different sets of articulators (body, face and hands) in this task.

After a short overview on the related work in Section 2, Section 3 introduces the corpus and Section 4 details the proposed methodology. The results are provided in Section 5.

2 Related Work

To our knowledge, this paper presents the first attempt of the task of automatic segmentation of SL into sentence-like units. This task has been suggested by Dreuw and Ney [8] as a tool for integration into a SL annotation program.

Despite a large amount of existing work for speech and text segmentation, there is debate surrounding the precise linguistic definition of a sentence in languages such as French or English [7]. Nevertheless, division by punctuation from written language is a good working solution for almost all cases. Automatic punctuation of speech can be achieved either using prosodic cues from audio or directly from a text transcription. On reference datasets, the former method tends to perform worse than the latter, but a combination of prosodic cues and a written transcription can have superior performance than either individually, as shown by Kolář and Lamel [13].

In SLs, purely oral languages, even a working notion of a sentence is unclear. Crasborn [6] proposes the pragmatic solution of identifying sentences in SL by firstly translating them into a written language and then calling a sentence the closest equivalent portion of SL to a sentence in the written language. This solution is somewhat unsatisfactory, as it requires translation to a written language. Nevertheless, Fenlon et. al. [9] demonstrate that both native signers and non-signers can reliably segment sentence boundaries in SL using visual cues such as head rotations, nodding, blinks, eye-brow movements, pauses and lowering the hands.

Our definition of a SU requires translation to a written language, but our goal is to learn to segment SL into sentence-like units purely from visual cues without translation into a written language. We note that SUs are not necessarily the same as what are sometimes called clauses, sentences or syntactic units in the linguistic literature on SL. Börstell et. al. [2] compare SUs with 'syntactic boundaries' annotated by a Deaf SL researcher. They find that many of the boundaries of the SUs overlap with the syntactic boundaries, but that there are more syntactic boundaries than there are SUs.

We consider SU boundary detection as a continuous SL recognition problem, as we learn visual cues in long sequences of video data. One main approach for continuous SL recognition consists of using RGB SL video as input, and then combining a 3D Convolutional Neural Network (CNN) with a Recursive Neural Network (RNN) to predict a sequence of words in the written language. Koller et. al. [14] use a CNN with a bi-directional LSTM (BiLSTM) and Huang et. al. [11] use a Hierachical Attention Network (HAN). Both of these articles use corpora in controlled environments with a single signer facing the camera.

Another main approach is to use sequences of skeleton data as input, which is arguably less dependent on the conditions of SL video production. Belissen et. al. [1] and Ko et. al. [12] use sequences of skeleton keypoints for continuous SL recognition, but concatenate the 2D skeleton keypoints into two vectors rather than exploiting the graph structure of the skeleton keypoints.

Yan et. al. [17] propose a Spatio-Temporal Graph Convolution Network (ST-GCN) for action recognition using sequences of skeleton keypoints that achieves state-of-the-art results. This model takes into account the spatio-temporal relationships between body keypoints. Our model is an adaptation of the ST-GCN, as this type of model is appropriate for our 2D skeleton video data. We combine the ST-GCN model with a BiLSTM, as we are predicting sequences not classes. This combination of a convolutional network and a BiLSTM is commonly used in language modelling [15].

3 Corpus

The MEDIAPI-SKEL corpus [4] contains 27h of subtitled French Sign Language (LSF) video in the form of sequences of 2D skeletons (see Fig. 1). This corpus has the unique quality of being both natural SL (produced outside laboratory conditions) and having accurately aligned subtitles. As far as we know, this is the only large existing corpus with these two characteristics.

The subtitles in this corpus are aligned to the SL video such that the video segment corresponds to the subtitle. The original language of almost all the videos is SL, which is then translated into written language for the subtitles.[3] The subtitles have been written by different people and aligned by hand, and so we expect some variation in the length and placement of the SUs.

[3] There are rare video segments where a hearing person is interviewed and this interview is translated into SL.

The 2D skeleton data contains 25 body keypoints, 2×21 hand keypoints and 70 facial keypoints for every person at every frame in the 27h hours of video content. Each 2-dimensional coordinate is also associated to a confidence value between 0 and 1.

This corpus contains 2.5 million frames associated to 20k subtitles, where each subtitle has an average length of 4.2 seconds and 10.9 words. The training data contains 278 videos, the validation data 40 videos, and the test data 50 videos. The average length of a video is 4.5 minutes. Videos may contain signers at different angles (not necessarily facing the camera) and around one-fifth of the videos contain multiple signers.

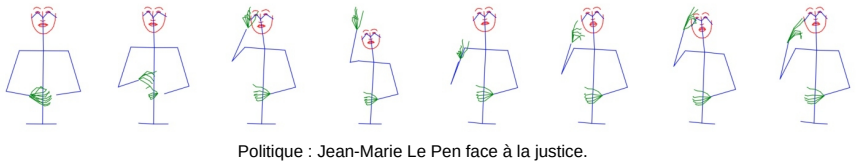

Fig. 1. MEDIAPI-SKEL corpus [4] with skeleton keypoints of LSF and aligned subtitles in written French. The graph structure connecting body keypoints (blue), face keypoints (red) and hand keypoints (green) is shown

Since the corpus contains dialogues between multiple people in various environments, it is necessary to clean the data automatically by detecting and tracking the current signer and by removing irrelevant keypoints.

The code for our skeleton data cleaning procedure and generation of labels is available online.[4] The main steps consist in:

- Converting all videos to 25 frames-per-second
- Omitting the legs and feet keypoints, as they are not relevant for SL, leaving us with a total of 125 keypoints
- Tracking each person in each video using a constraint on the distance between body keypoints between consecutive frames
- Omitting people unlikely to be signers, specifically those with hands outside of the video frame, those with hands that hardly move, those that are too small (in the background of the video) or those that appear only for very short time periods (under 10 frames)
- In the case of multiple potential signers, choosing the most likely signer in each second of video based on a criterion involving hand size times variation of wrist movement of the dominant hand
- Imputation of missing skeleton keypoints using past or future frames
- Temporal smoothing with a Savitzky-Golay filter

[4] `https://github.com/hannahbull/clean_op_data_sl`

Our final input data consist of temporal sequences of variable lengths of 2D skeleton keypoints corresponding to individuals in SL video.

We label a frame of a sequence with 0 if there is no subtitle associated to that frame or if the frame is within a distance of 2 frames from a frame with no associated subtitle. We label all other frames as 1. The padding of the 0-labelled frames partially controls for the fact that the SUs are not precise at the frame-level. Frames labelled 1 are SUs, and frames labelled 0 are SU boundaries.

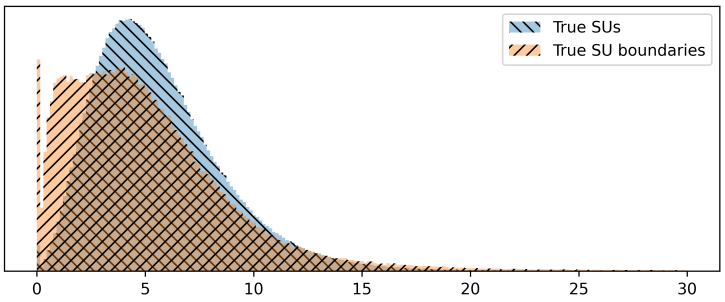

Fig. 2. Density histogram of the average velocity of the 15 upper body keypoints of likely signers in the training set. Units are pixel distance moved per frame with 1080p resolution

Fig. 2 shows the distribution of the average velocity of the body keypoints of likely signers in the training set by label. Sequences where there is unlikely to be a signer due to lack of hand visibility or hand movement are omitted using our data cleaning procedure. True SU boundaries tend to have lower average body keypoint velocity compared to true SUs, but velocity is an insufficient indicator to predict SU boundaries in SL discourse.

4 Methodology

4.1 Model

Our model is a spatio-temporal graph convolutional network (ST-GCN) following Yan et. al. [17], which we adjoin to a BiLSTM network to capture the sequential nature of the output (Fig. 3). The spatial graph structure of the body keypoints, face keypoints and hand keypoints follows the human joint structure. The temporal graph structure connects body keypoints across time. The edge importance in the graph is learned during training. The convolution operation is across the spatial and temporal edges of the graph.

The ST-GCN architecture is identical to that used by Yan et. al. [17], but without temporal pooling. The model is composed of 9 layers of ST-GCN units,

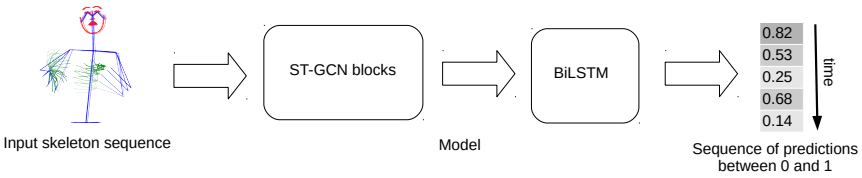

Fig. 3. ST-GCN+BiLSTM model on skeleton sequence for SU detection

where the first 3 layers have 64 output units, the second 3 layers have 128 output units and the final 3 layers have 256 output units. The embedding dimension of the BiLSTM is thus 256 and we also set the hidden dimension of the BiLSTM to be 256.

Each input sequence of skeleton keypoints has a length of 125 frames, but we take every second frame of the video, so this corresponds to a sequence length of 10s. This means that we expect around two or three SUs per sequence, as the average subtitle length is 4.2s.

Each skeleton sequence is normalised such that the mean and variance of the x-coordinates and y-coordinates of the skeleton over time are equal to 0 and 1. During training, we add random flips to the horizontal dimension of the skeleton keypoints in order to take into account for left-handed and right-handed signers. We also shuffle the order of skeleton sequences at each epoch.

We use SGD optimisation with a learning rate of 0.01, a weight decay of 0.0001, Nesterov momentum of 0.9 and binary cross-entropy loss. The model is trained for 30 epochs. Due to memory constraints, the batch-size is 4.

4.2 Experiments

We train our model on 278 videos and test our model on 50 videos. Our full model uses 15 body keypoints, 70 face keypoints and 2×21 hand keypoints shown respectively in blue, red and green in Fig. 1. In order to understand the contributions of the body, face and hand keypoints, we train the model using only the body keypoints, only the face keypoints and only the hand keypoints, as well as the body + face, the body + hand and the body + face + hand keypoints. We keep the architecture of the model constant.

Moreover, we compare the performance of our model between videos with one signer and videos with multiple signers. The videos with multiple signers often contain dialogues between people not necessarily facing directly at the camera. This is to test the robustness of our model to more diverse scenarios.

4.3 Evaluation Criteria

Our evaluation metrics should take into account that SUs are not annotated by the subtitler at a frame-level accuracy. We propose both frame-wise and unit-wise metrics, allowing for shifts in SUs.

As a flexible frame-wise metric, we propose dynamic time warping (DTW) with a window constraint as an evaluation criteria. This computes the distance between the true sequence and the predicted sequence of SUs, allowing for frames to be shifted within a certain window length w. We compute this DTW accuracy for different values of the window length w. When $w = 0$, this is the frame-wise difference between the predicted SUs and the true SUs. We also compute the DTW distance for $w \in \{5, 10, 15\}$, which corresponds to the minimum frame-wise difference between the predicted SUs and the true SUs allowing for frames to be shifted up to 5, 10 or 15 frames.

Additionally, we compute the ROC-AUC statistic, the frame-wise precision, recall and $F1$-score. The precision is given by the number of frames correctly identified with the label 0 divided by the total number of frames identified with the label 0. The recall is given by the number of frames correctly identified with the label 0 divided by the total number of true frames with the label 0. The $F1$ score is the harmonic mean of precision and recall.

Furthermore, we consider unit-wise evaluation metrics, allowing for 15 frame (0.6s) shifts in SU boundaries. We match each predicted SU boundary to the closest true SU boundary, where the closest true SU boundary is defined as the true SU boundary with the greatest intersection with the predicted SU boundary, or, in the case of no intersection, the closest true SU boundary within 15 frames. Calculating the number of matches divided by the total number of predicted SU boundaries gives us a unit-wise precision metric. In the same way, we can match each true SU boundary to the closest predicted SU boundary. The number of matches divided by the total number of true SU boundaries gives us a unit-wise recall metric. From this precision and recall metric, we can compute a unit-wise $F1$ score.

5 Results and Discussion

Table 1 shows frame-wise evaluation metrics on the test set. Our results are encouraging and we obtain a ROC-AUC statistic of 0.87 for our predictions, with the highest score obtained using the body, face and hand keypoints. Instead of relying on the frame-wise error rate, it is important to account for slight shifts in SUs as those who subtitle the videos do not aim for accuracy at the level of the frame. Allowing for shifts of up to 0.6s (15 frames), we obtain a frame-wise error rate of 8% when using only the body keypoints. Table 2 presents unit-wise evaluation results and shows that 76% of true SU boundaries can be associated to a predicted SU boundary within 15 frames.

When asking native signers to annotate sentence boundaries in SL, Fenlon et. al. [9] found inter-participant agreement of sentence boundary annotation

within 1 second to be around 63%. Whilst this is not exactly the same task as subtitling SL video, we can expect that there is quite a high degree of variation in the choice of subtitle boundaries. In light of this finding, our error rate seems reasonable.

Table 1. Frame-wise evaluation metrics on the test set. The full model uses face, body and hand keypoints. The pre-processing version shows an evaluation after annotation of segments without an identified signer as not belonging to SUs. The final line shows the results for a constant prediction. DTW0 is the frame-wise prediction error. DTW5, DTW10 and DTW15 are the DTW errors respectively allowing for a 5, 10 and 15 frame discrepancy in predictions

	DTW0	**DTW5**	**DTW10**	**DTW15**	**AUC**	**Prec.**	**Recall**	**F1**
full	0.1660	0.1255	0.1045	0.0927	**0.8723**	0.5023	**0.7510**	0.6019
face+body	0.1560	0.1172	0.0973	0.0868	0.8708	0.5241	0.7259	0.6087
body+hands	0.1661	0.1269	0.1064	0.0952	0.8659	0.5023	0.7380	0.5977
face	0.1858	0.1483	0.1248	0.1100	0.8325	0.4624	0.6830	0.5514
body	**0.1410**	**0.1055**	**0.0882**	**0.0790**	0.8704	**0.5616**	0.7122	**0.6280**
hands	0.1821	0.1417	0.1186	0.1053	0.8554	0.4713	0.7360	0.5747
Pre-processing	*0.1406*	*0.1365*	*0.1333*	*0.1309*	*0.6039*	*0.7828*	*0.2201*	*0.3436*
Constant pred.	*0.1672*	*0.1672*	*0.1672*	*0.1672*	*0.5000*	*0.1671*	*1.0000*	*0.2865*

Table 2. Unit-wise evaluation metrics on the test set allowing for 15 frame (0.6s) shifts in SU boundaries. The full model uses face, body and hand keypoints. The pre-processing version shows an evaluation after annotation of segments without an identified signer as not belonging to SUs

	Prec.	**Recall**	**F1**
full	0.6609	**0.7631**	0.7083
face+body	0.6840	0.7408	**0.7113**
body+hands	0.6250	0.7492	0.6815
face	0.6403	0.6909	0.6646
body	**0.7090**	0.6866	0.6976
hands	0.6147	0.7619	0.6804
Pre-processing	*0.9341*	*0.0803*	*0.1478*

Part of the accuracy of our model is accounted for by pre-processing the data to label obvious SU boundaries, such as moments where there are no signers in the video. Such frames are correctly identified as having no associated subtitle 78% of the time, as noted in the second last line of Table 1. Errors here seem to be mostly due to subtitles extending beyond scenes containing signers, rather than failure to detect a signer in a scene, however further annotation of signers

would be needed to verify this. Our ST-GCN+BiLSTM model makes significant improvements on top of this pre-processing.

From Table 1, we see that the full model has the highest ROC-AUC statistic and the highest recall, suggesting that including the facial and hand keypoints detects the most SU boundaries. However, the body model makes fewer incorrect predictions of SU boundaries and has a higher precision. Our unit-wise metrics in Table 2 reinforce this observation. The full model correctly identifies 76% of the true SU boundaries within 15 frames, but the body model has the highest precision with 71% of the predicted SU boundaries within 15 frames of a true SU boundary. Börstell et. al. [2] find that there are more 'syntactic boundaries' than SUs. Perhaps our full model is good at learning visual cues of such 'syntactic boundaries', which do not always correspond to actual SU boundaries.

Fig. 4 shows an example of the predictions and true labels on a video from the test set using the full model. Most of the true SU boundaries are correctly detected, however there is an over-detection of SU boundaries. Fig. 7 shows that the predicted lengths of SUs using the full model is shorter than the true lengths of SUs. This difference in length is less pronounced when using the body model.

Morover, predicted SU boundaries tend to be slightly longer than the true SU boundaries. The median difference between predicted SU boundaries and the associated true SU boundaries within 15 frames is around 5-7 frames in all our models. The median absolute difference between predicted SU boundaries and the associated true SU boundaries is 7-9 frames. The problem of over-detection or under-dectection of SU boundaries and differences in lengths could be alleviated by assigning length and regularity priors to the SUs. This is similar to applying shape priors in image segmentation [5], [16].

Fig. 5 and Fig. 6 show examples of correct and incorrect predictions from Fig. 4. The left of Fig. 5 shows an example of an obvious SU boundary where the signer pauses with their hands folded. This is correctly predicted by our model, albeit our predicted SU boundary is a little longer than the true boundary. The right of Fig. 5 shows a SU boundary with more subtle visual cues, including the head turning towards the camera and a slight deceleration of movement. This is also correctly detected by our model, but with a slight shift of about half of a second.

The left of Fig. 6 shows an SU boundary detected by our model but which is not a true SU boundary. However, this particular example could have been an SU boundary had the subtitles for this video been aligned differently. Some of our incorrectly detected SU boundaries are thus likely to correspond to sentence-like boundaries but which are simply not annotated as such by the subtitler. The right of Fig. 6 shows a SU boundary not detected by our model. This particular SU boundary does not have clear visual cues, and its detection may perhaps require an understanding of the SL sequence.

Facial visual cues for semantic boundaries in SL can include blinks, eyebrow movements, head nodding or turning the head to stare directly at the camera. Manual cues include specific hand movements and the signer folding their hands together at the waist level. We thus assess whether or not including facial and

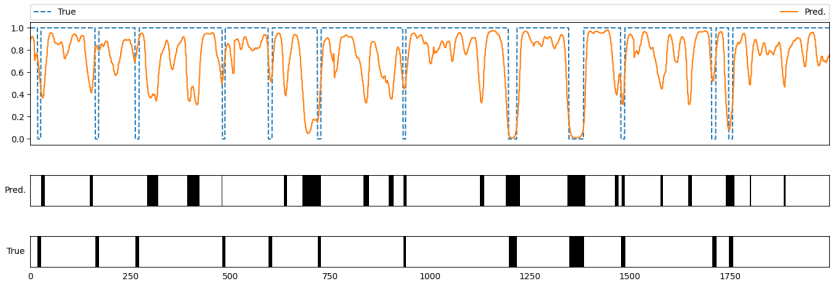

Fig. 4. True and predicted labels for a video sequence using the full (face+body+hands) model

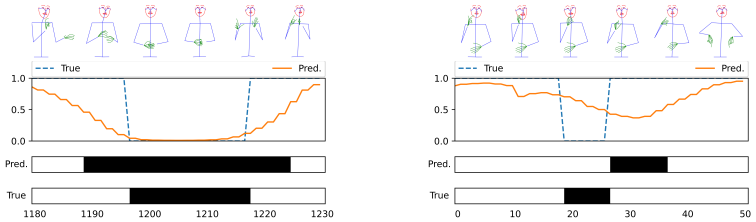

Fig. 5. Correctly detected SU boundaries from Fig. 4

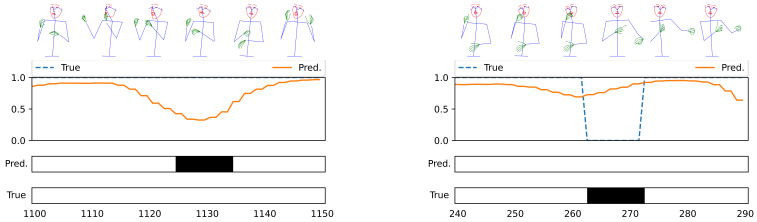

Fig. 6. Incorrectly detected SU boundaries from Fig. 4

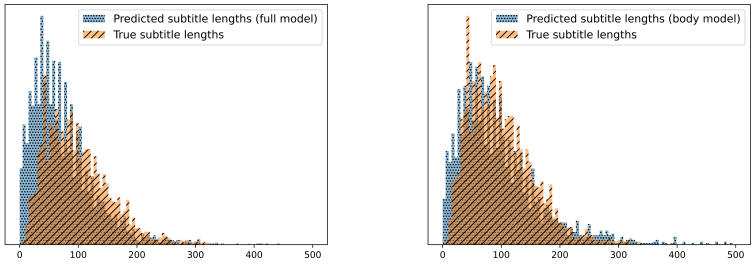

Fig. 7. Length of true SUs compared to predicted SUs

hand keypoints improves SU detection. We cannot conclude that adding the face and the hand keypoints to the body model makes a significant improvement to SU detection. Nevertheless, the face keypoints or the hand keypoints alone make surprisingly accurate predictions. The face model has a ROC-AUC statistic of 0.83. Subtle facial cues are likely to be picked up by our model. Similarly, the hands alone make relatively accurate predictions.

Table 3. Evaluation metrics for videos with one signer and videos with multiple signers. Models and evaluation metrics are as in Table 1

	DTW0	**DTW5**	**DTW10**	**DTW15**	**AUC**	**Prec.**	**Recall**	**F1**
full 1 signer	0.1366	0.0959	0.0776	0.0686	**0.8876**	0.5144	**0.7456**	0.6088
body 1 s.	**0.1204**	**0.0838**	**0.0676**	**0.0601**	0.8854	**0.5611**	0.7140	**0.6284**
full >1 s.	0.2227	0.1824	0.1562	0.1392	**0.8388**	0.4878	**0.7579**	0.5936
body >1 s.	**0.1809**	**0.1474**	**0.1278**	**0.1156**	0.8365	**0.5622**	0.7100	**0.6275**

As shown in Table 3, accuracy is reduced amongst test videos with more than one signer, but the ROC-AUC statistic is still relatively high at 0.84. The DTW error rate with a window length of 15 is 12%. On videos in the test set with one signer, the ROC-AUC statistic is 0.89 and the DTW error rate with a window length of 15 frames is only 6%. This suggests that our model is robust to natural SL video, including examples of dialogue between multiple signers.

6 Conclusion

We provide baseline results for automatic segmentation of SL video into sentence-like units. We use natural SL video and allow multiple signers and camera angles. Our results are encouraging, given the variability of identification of semantic boundaries in a SL discourse across different annotators and given the fact that the SU annotations are not accurate at the frame level.

Our full model using face, body and hand keypoints has a high recall statistic but finds more SUs than necessary. We are interested to find out whether or not these additional SU boundaries correspond to semantic boundaries in the SL discourse that are not annotated by the subtitler. Further annotation of our test data would be required in order to see whether or not this is the case.

We would like to improve our model by better controlling the final distribution of the SUs. For example, we would like to be able to set priors on the duration of the SUs in order to control the length of segments and the regularity of the segmentation.

Due to the relative lack of understanding of SL grammar and the lack of a written form of SL, we are constrained to the detection of prosodic cues for this segmentation. In future work, we intend to improve detection of SUs by additionally identifying certain signs.

Acknowledgements

This work has been partially funded by the ROSETTA project, financed by the French Public Investment Bank (Bpifrance). Additionally, we thank *Média-Pi !* for providing the data and for the useful discussions on this idea.

References

1. Belissen, V., Braffort, A., Gouiffès, M.: Dicta-Sign-LSF-v2: Remake of a continuous french sign language dialogue corpus and a first baseline for automatic sign language processing. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20). pp. 6040–6048. European Language Resource Association (ELRA), Marseille, France (May 2020)
2. Börstell, C., Mesch, J., Wallin, L.: Segmenting the swedish sign language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In: Beyond the Manual Channel. Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages. pp. 7–10 (2014)
3. Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al.: Sign language recognition, generation, and translation: An interdisciplinary perspective. In: The 21st International ACM SIGACCESS Conference on Computers and Accessibility. pp. 16–31 (2019)
4. Bull, H., Braffort, A., Gouiffès, M.: MEDIAPI-SKEL - a 2D-skeleton video database of french sign language with aligned french subtitles. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20). pp. 6063–6068. European Language Resource Association (ELRA), Marseille, France (May 2020)
5. Chan, T., Zhu, W.: Level set based shape prior segmentation. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2, pp. 1164–1170. IEEE (2005)
6. Crasborn, O.A.: How to recognise a sentence when you see one. Sign Language & Linguistics **10**(2), 103–111 (2007)
7. De Beaugrande, R.: Sentence first, verdict afterwards: On the remarkable career of the sentence. Word **50**(1), 1–31 (1999)
8. Dreuw, P., Ney, H.: Towards automatic sign language annotation for the ELAN tool. In: Proceedings of the Third LREC Workshop on Representation and Processing of Sign Languages. pp. 50–53. European Language Resource Association (ELRA), Marrakech, Morocco (May 2008)
9. Fenlon, J., Denmark, T., Campbell, R., Woll, B.: Seeing sentence boundaries. Sign Language & Linguistics **10**(2), 177–200 (2007)
10. Filhol, M., Hadjadj, M.N., Choisier, A.: Non-manual features: the right to indifference. In: 6th Workshop on the Representation and Processing of Sign Languages: Beyond the manual channel. Satellite Workshop to the 9th International Conference on Language Resources and Evaluation (LREC 2014). pp. 49–54 (2014)
11. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
12. Ko, S.K., Kim, C.J., Jung, H., Cho, C.: Neural sign language translation based on human keypoint estimation. Applied Sciences **9**(13), 2683 (2019)
13. Kolář, J., Lamel, L.: Development and evaluation of automatic punctuation for french and english speech-to-text. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
14. Koller, O., Zargaran, S., Ney, H.: Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3416–3424 (July 2017). https://doi.org/10.1109/CVPR.2017.364

15. Sundermeyer, M., Ney, H., Schlüter, R.: From feedforward to recurrent lstm neural networks for language modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(3), 517–529 (2015)
16. Veksler, O.: Star shape prior for graph-cut image segmentation. In: European Conference on Computer Vision. pp. 454–467. Springer (2008)
17. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)