Context Matters: Self-Attention for Sign Language Recognition

Fares Ben Slimane and Mohamed Bouguessa

University of Quebec at Montreal

Abstract

This paper proposes an attentional network for the task of Continuous Sign Language Recognition. The proposed approach exploits co-independent streams of data to model the sign language modalities. These different channels of information can share a complex temporal structure between each other. For that reason, we apply attention to synchronize and help capture entangled dependencies between the different sign language components.

Contributions

- Devising an end-to-end framework for sequence to sequence Sign Language Recognition that utilizes self-attention for temporal modeling.
- Elaborating a more efficient method to incorporate handshapes with their spatiotemporal context for Sign Language Recognition.
- Achieving competitive results, in terms of Word Error Rate, on the RWTH-PHOENIX-Weather 2014 benchmark dataset.

Introduction

Sign languages are often defined as manual languages. However, besides the hand articulations, non-manual components like facial expressions, arm, head, body movements, and positions play a crucial part in Sign Languages. Any change in one of these components can alter the meaning of a sign. Usually, the handshape performed by the dominant hand carries most of the meaning of the sign. Accordingly, in this paper, we propose an attention-based approach for sequence to sequence sign language alignment and recognition. Unlike previous works, the originality of our approach lies in explicitly picking up and aggregating contextual information from the non-manual sign language components. Without any domain annotation, our approach is able to exclusively identify the most relevant features associated with the handshape when predicting a sign.

Proposed Approach

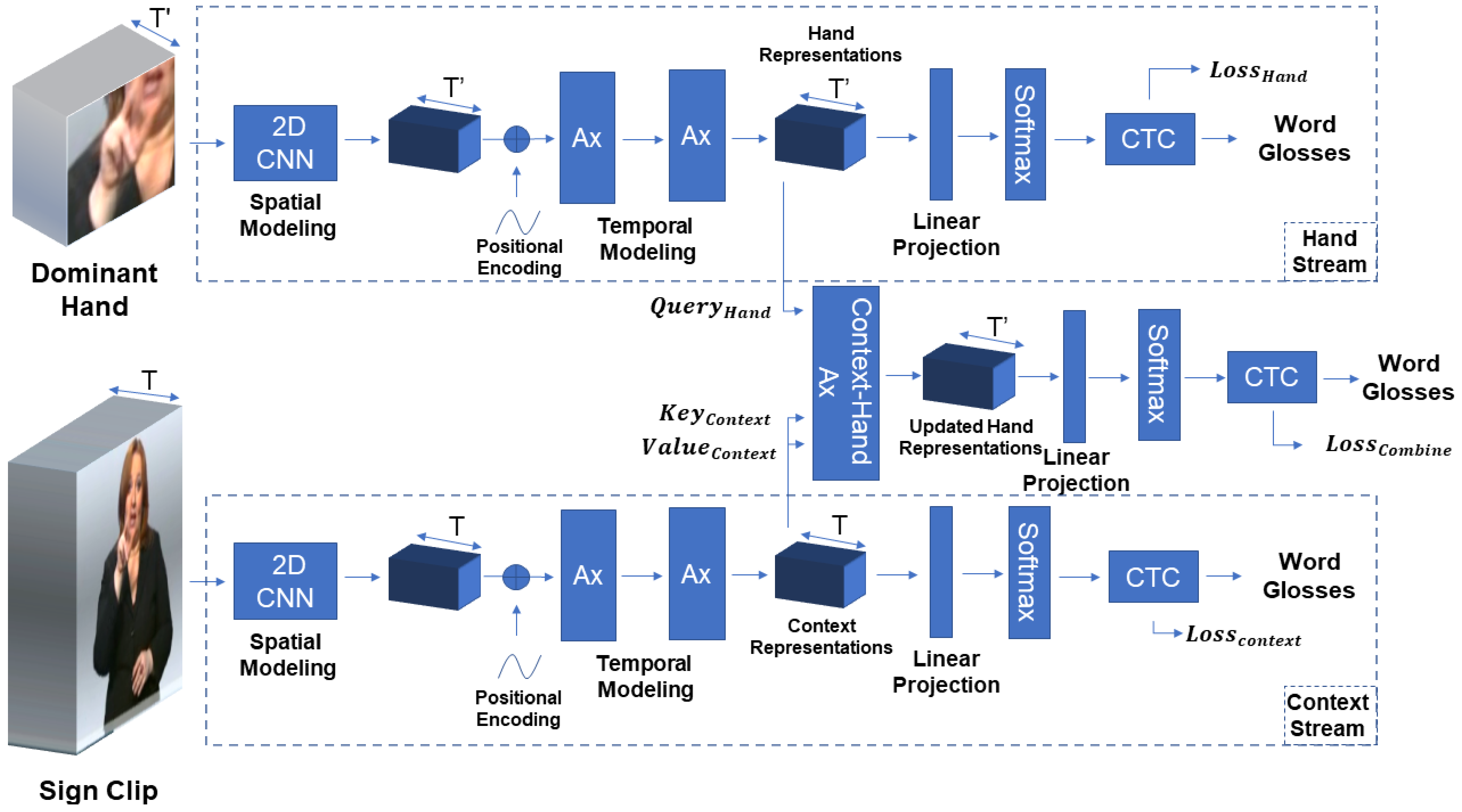
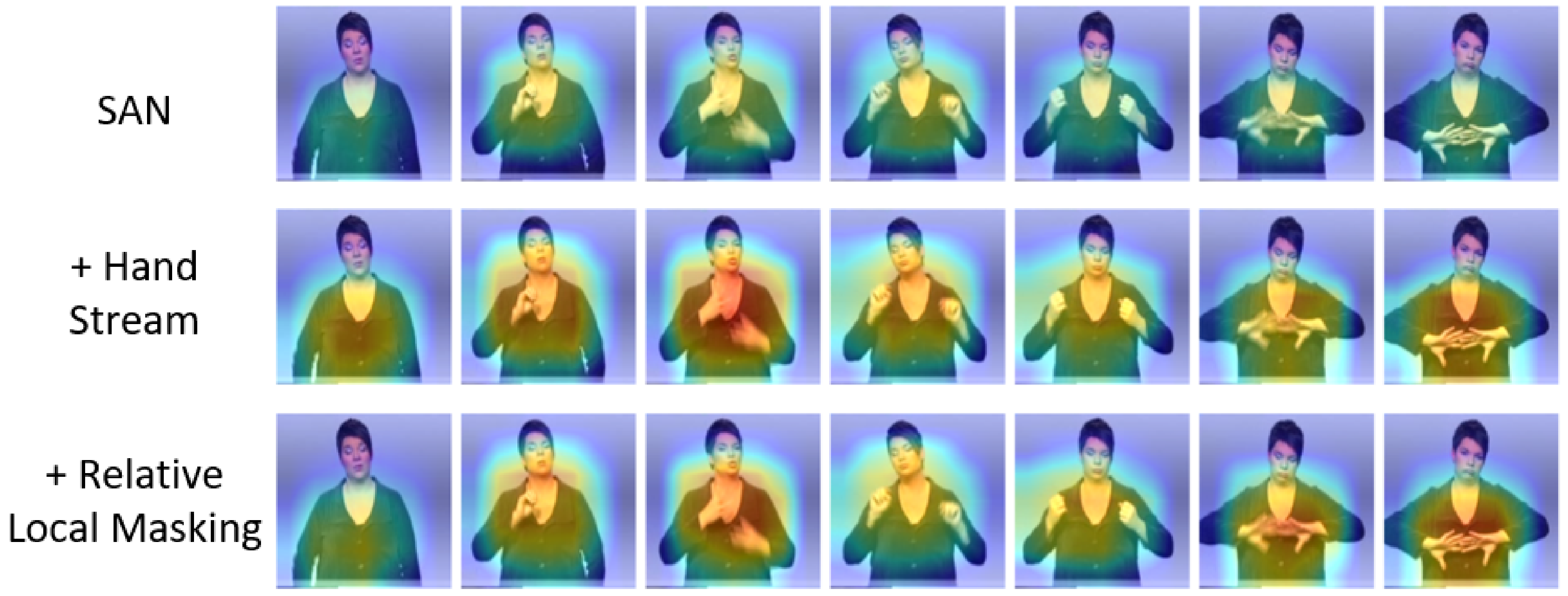


Figure 1:Combination of both the full-frame and the handshape streams through a Context-Hand Attention layer.

Approach Variations

- First, Sign Attention Network (SAN) that employs self-attention for temporal modeling.
- Second, we add a secondary stream for the dominant handshape sequences and we combine the hand features with their Spatio-temporal full-body context.
- Third, instead of considering the entire context information, we merely attend to information from the handshape local surroundings by applying a local relative mask. This will allow the model to only focus on the required context, discarding unnecessary distant information.

Qualitative Analysis



SAN with hand stream primarily focus on the dominant hand (right hand) and the face area which reinforces the intuition that our model is able to **identify the essential components for sign interpretations**.

Quantitative Results

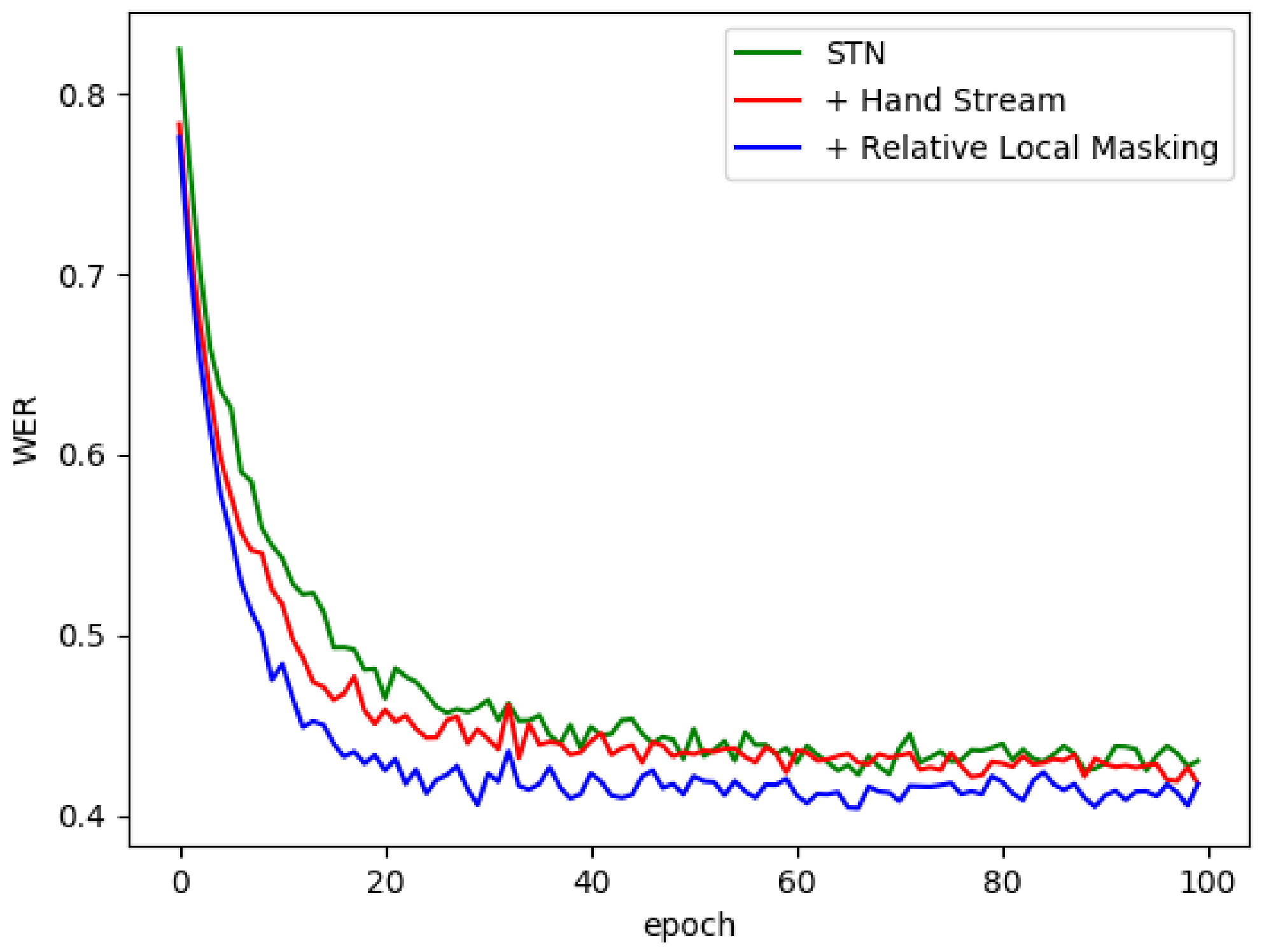


Figure 2:The Word Error Rate learning curve of our three SAN variants.

As shown in Figure 2, Adding handshape features improves training and accelerates model convergence. This empirically showcases the usefulness of combining the dominant hand with the overall context derived from the nonmanual components of the sign.

	Dev	Test
SAN	35.33	35.45
+ Hand Stream	33.68	34.12
+ Relative Local Masking	**32.74**	**33.29**

Conclusion

In this work, we have proposed a novel method that exploits attention to effectively combines hand query features with their respective temporal full-body context without the need for any additional supervision. We have proven the efficiency of such an approach to the task of Continuous Sign Language Recognition.

Acknowledgements

This work is supported by research grants from the Natural Sciences and Engineering Research Council of Canada(NSERC).