# Sign Language Translation using Motion Filters and Attention Models

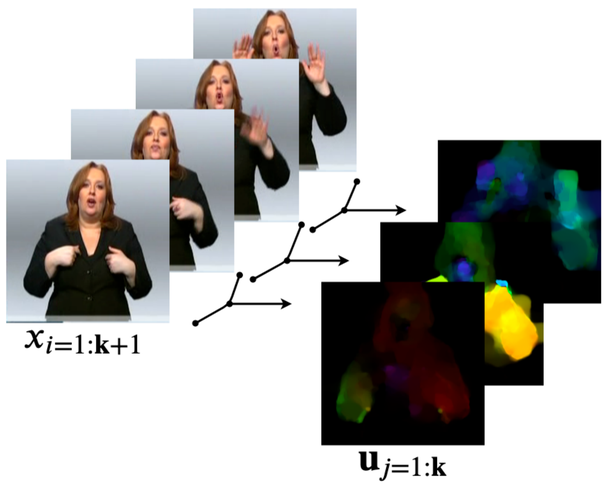
Jefferson Rodríguez and Fabio Martínez
jefferson.rodriguez2@saber.uis.edu.co, famarcar@saber.uis.edu.co
Biomedical Imaging, Vision and Learning Laboratory - BIVL$^2$ab
Motion Analysis and Computer Vision - MACV
Universidad Industrial de Santander - Bucaramanga, Colombia

## SL Translation

More than 70 million people use at least one Sign Language (SL) as their main channel of communication. Translating a video sign language into a spoken language is a complex task as the structure and grammar of each language must be processed at the same time.

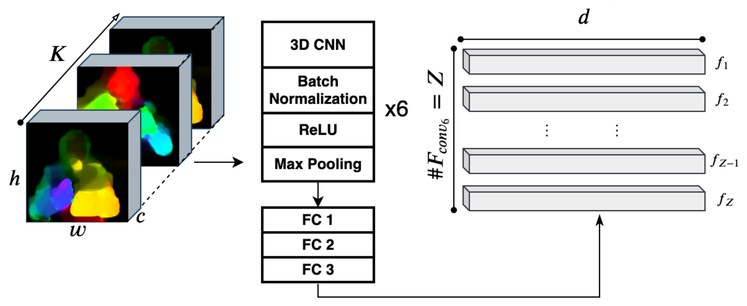
## Motion Information

Signs are articulated motions with a temporal coherence. Therefore, motion is a fundamental component of SL, which determines transitions between gestures and contributes to the grammatical structure of language. In this work was used a dense optical flow that considers large coherent motion displacements and long-term dependencies during the sequence [1].




## Video Representation

Inspired by [2], a video-level sign representation module was implemented that captures the main dynamic patterns throughout the temporal sequence and maintains the most relevant spatial information. This module is composed mainly of six 3D convolutional layers followed by three fully connected dense layers. Then, the architecture takes the $\mathbf{U}$ volumes and returns a motion convolved representation matrix $\mathbf{F} = \{f_1, \dots f_Z\}, f_z \in \mathbb{R}^d$, that represent the $d-$dimensional $Z$ filters in the last CNN layer.



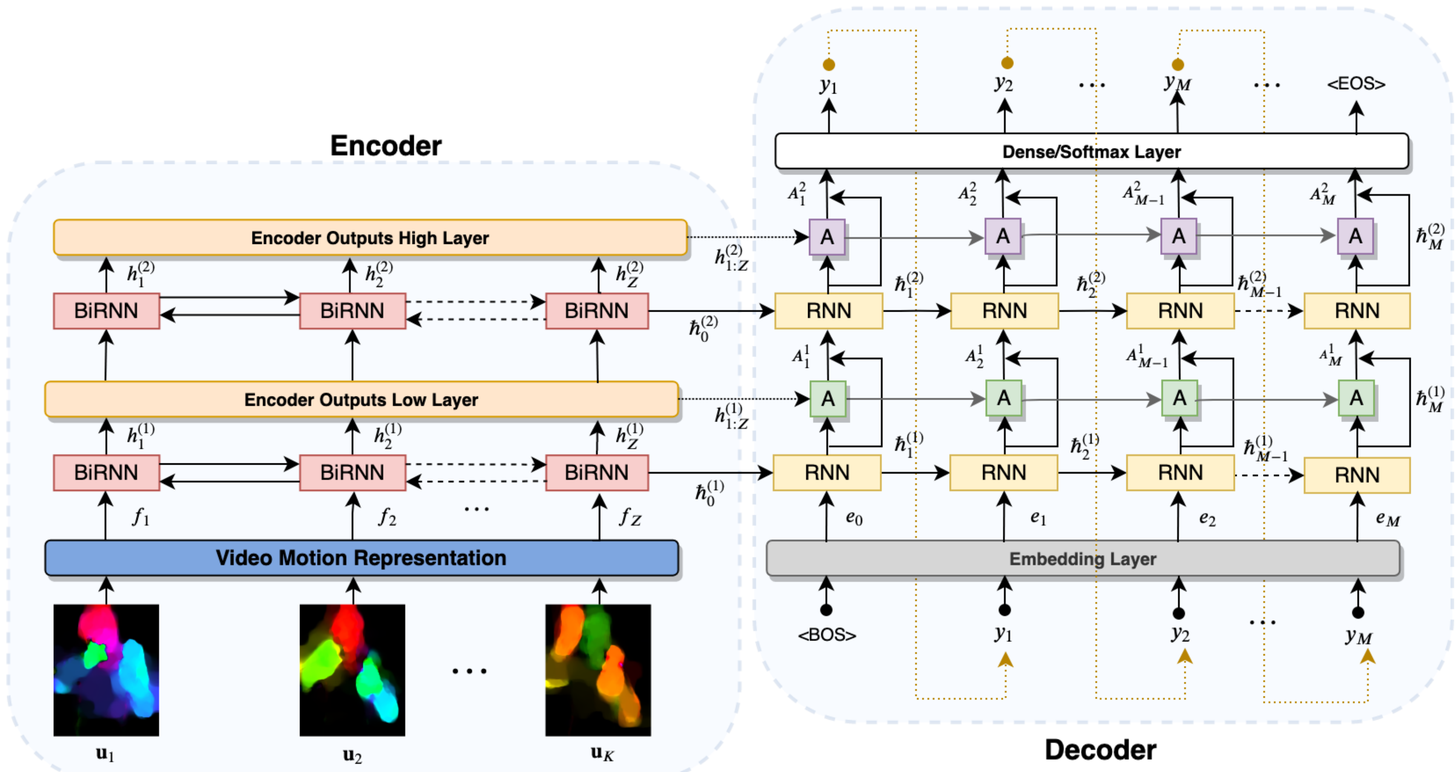

## Encoder-Decoder Net

The translation of a video sign $\mathbf{X} = (x_1, x_2, \dots, x_{k+1})$ with $\mathbf{k}+1$ frames in a sequence of text $\mathbf{y} = (y_1, y_2, \dots, y_M)$ with $M$ words, can be done by an encoder-decoder architecture directly modeling the conditional probability $p(\mathbf{y}|\mathbf{X})$. The encoder is a two-layer ($l = 2$) bi-directional recurrent network that captures temporal dependencies between motion descriptors $\mathbf{F}$ in states $[\overrightarrow{h}^{(l)}_{1:Z}; \overleftarrow{h}^{(l)}_{Z:1}]$. These historical kinematic states are used by the decoder to decompose the joint probability into ordered conditional probabilities:

$$p(\mathbf{y}|\mathbf{X}) = \prod_{m=1}^{M} p(y_m|y_1, y_2, \dots, y_{m-1}, (h_Z^{(1)}, h_Z^{(2)}))$$

This conditional probability is solved by implementing an unidirectional RNN with motion attention mechanism $A_m^l$ to relate both language modes. The states $(h_Z^1, h_Z^2)$ are updated with the states $(\hbar_m^{(1)}, \hbar_m^{(2)})$ generated in the decoder.

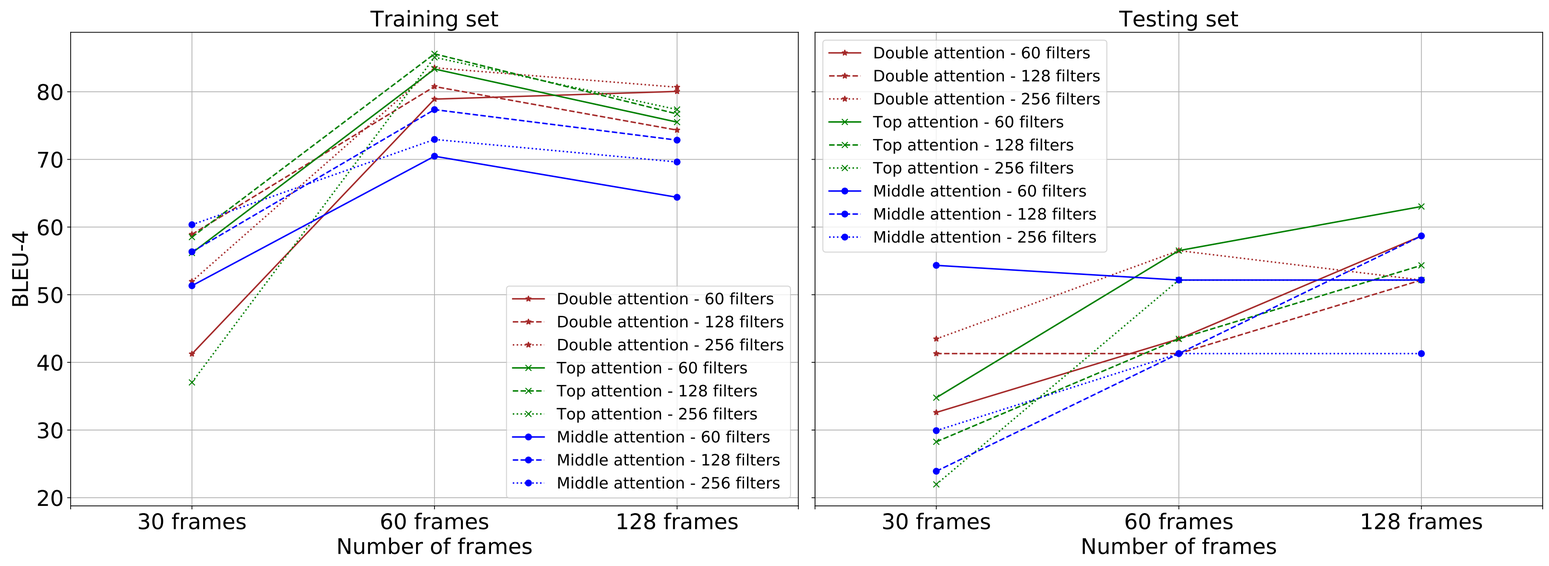
## Architecture for SL translation




**General encoder-decoder architecture with attention of the proposed translation approach.** The system is composed of three important modules that analyze motion patterns at different levels of complexity: Convolutional representation of motion, coding of temporal patterns and their correlation with sentences in spoken language.

## Results

In this approach, is important the number of filters computed in the last CNN layer, as well as the number of frames selected and the position of the attention. For this, the corresponding experimentation was carried out to determine the effect of each parameter on the translation using motion information.




**a)** Shows the results contrasting the performance in the two types of information in our Colombian SL dataset with ∼1600 video signs clips on a 53-word vocabulary for BLEU 4-gram score.

**b)** Shows the performance obtained in RWTH-PHOENIX German dataset [3] with ∼7100 video signs on a ∼2900-word vocabulary using the best setups and BLEU 4-gram score.

(a) CSL Dataset

| a. | Train | Test | Setup |
|---|---|---|---|
| *Double Attention (Top and Middle)* | | | |
| Flow | 74.38 | 52.17 | 128K - 128Z |
| RGB | 16.04 | 19.56 | 128K - 128Z |
| *Single Attention (Top)* | | | |
| Flow | **76.77** | 54.34 | 128K - 128Z |
| RGB | 21.64 | 13.98 | 128K - 128Z |
| *Single Attention (Middle)* | | | |
| Flow | 72.90 | **58.69** | 128K - 128Z |
| RGB | 8.29 | 0.0 | 128K - 128Z |

(b) Phoenix Dataset

| | Dev | Test | Setup |
|---|---|---|---|
| *Double Attention (Top and Middle)* | | | |
| Flow | 2.41 | 3.08 | 60K - 256Z |
| RGB | 0.56 | 0.56 | 60K - 256Z |
| *Single Attention (Top)* | | | |
| Flow | 3.82 | 3.73 | 60K - 60Z |
| RGB | **4.68** | **5.32** | 60K - 60Z |
| *Single Attention (Middle)* | | | |
| Flow | 1.7 | 2.35 | 30K - 60Z |
| RGB | 4.12 | 4.81 | 30K - 60Z |

## Conclusions

- A compact recurrent encoder-decoder network (9M parameters) that hierarchically exploits motion and temporal information for SL translation was proposed.
- The obtained results confirm the advantage of motion-based representations face to typical appearance in complex and small datasets.
- The results show that the best position of the attention is at the top level of the decoder.

## References

[1] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2010.

[2] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. *CVPR 2018 Proceedings*, 2018.