DYNAMIC SIGN LANGUAGE RECOGNITION COMBINING DYNAMIC IMAGES AND CONVOLUTIONAL NEURAL NETWORK

LatinX in AI Research at ICML 2020

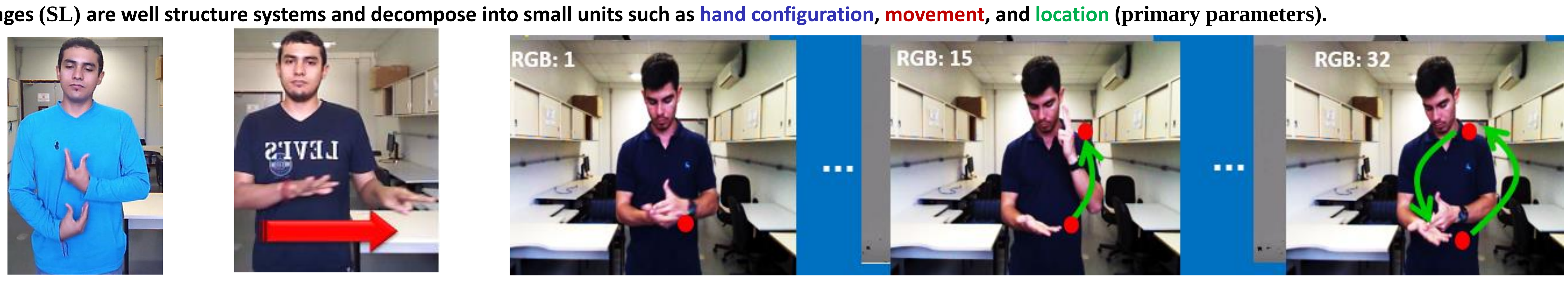
Lourdes Ramirez Cerna*, Edwin Escobedo Cardenas+

*Department of Informatics, National University of Trujillo in Peru -- UNT

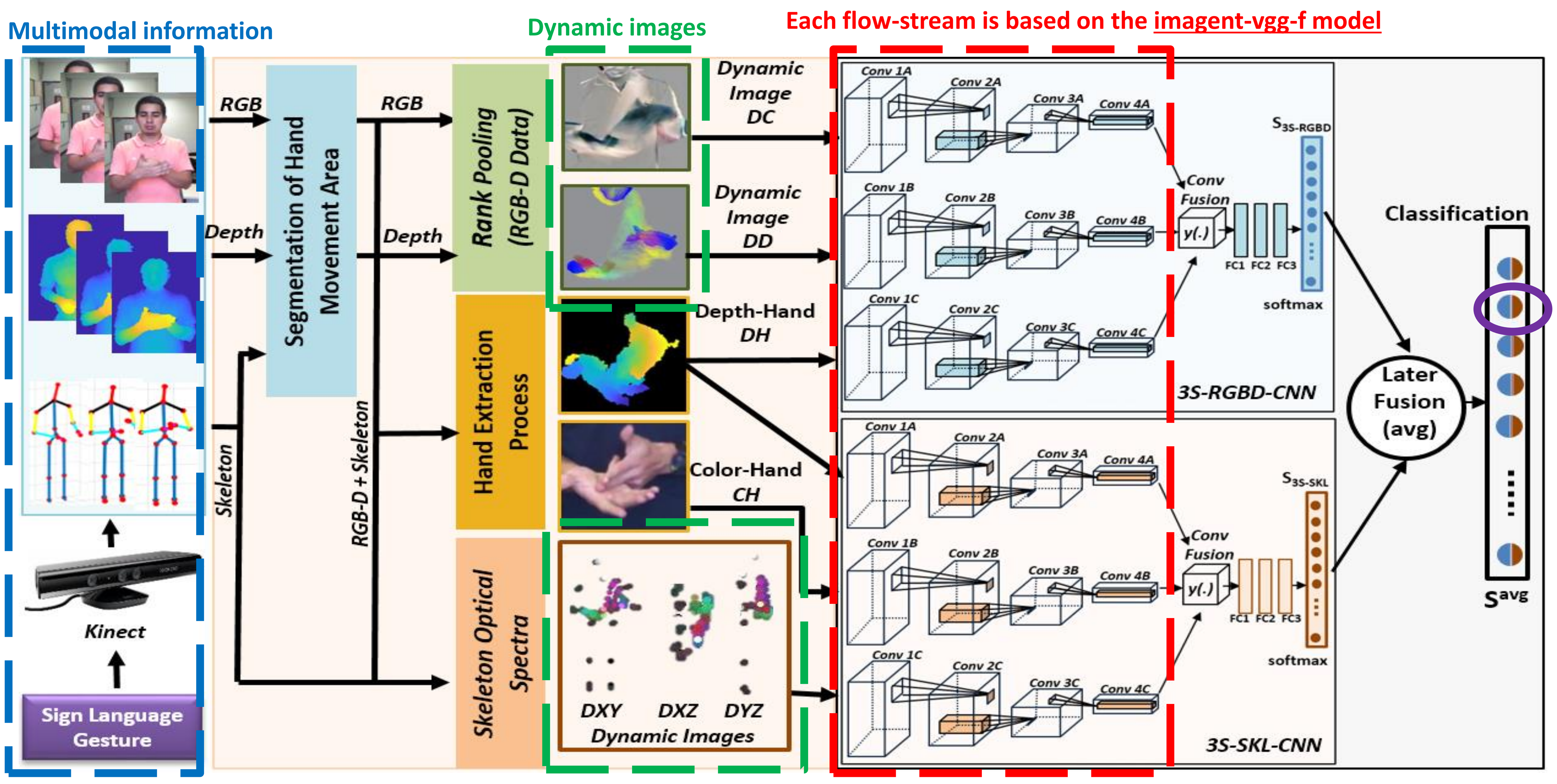
+ Department of computer Science, Federal University of Ouro Preto in Brazil -- UFOP

1. INTRODUCTION

Sign languages (SL) are well structure systems and decompose into small units such as hand configuration, movement, and location (primary parameters).

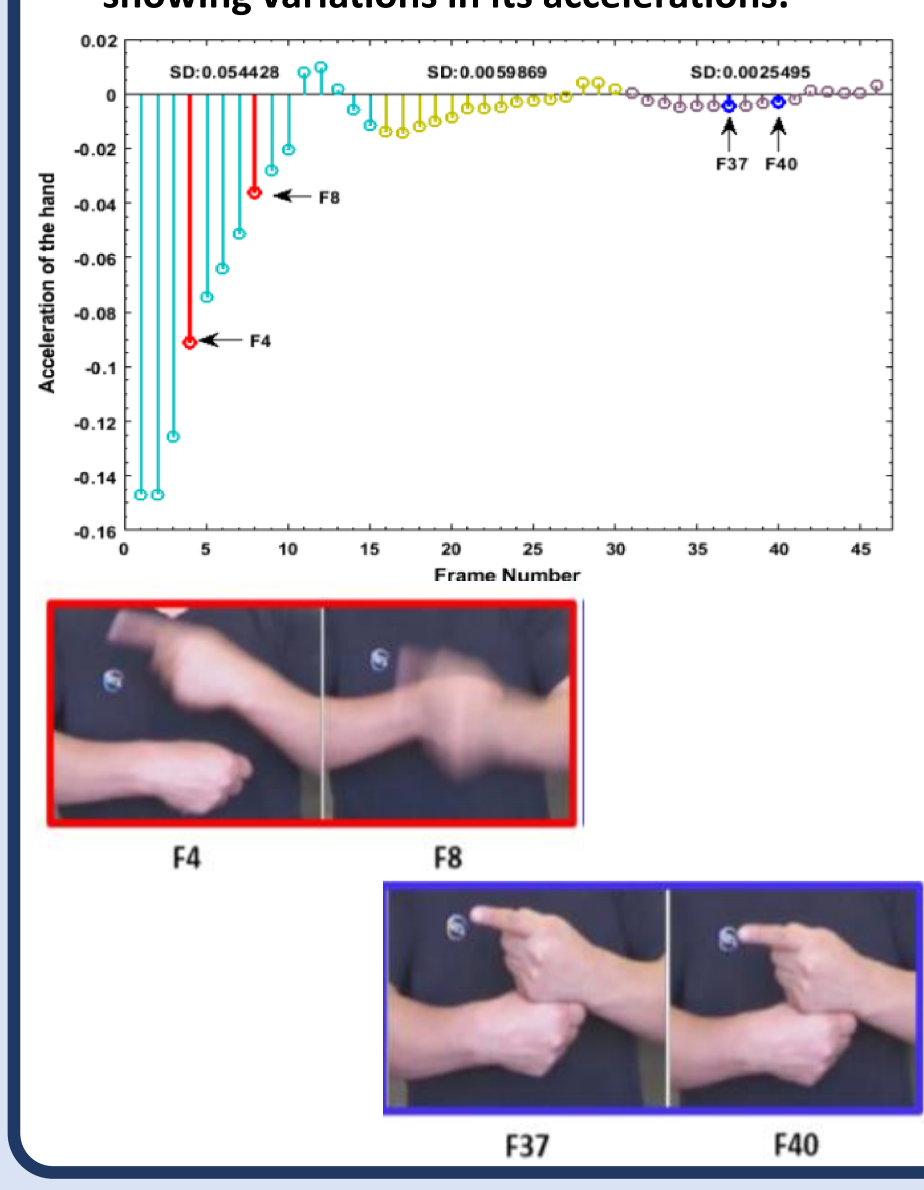


2. PROPOSED METHOD



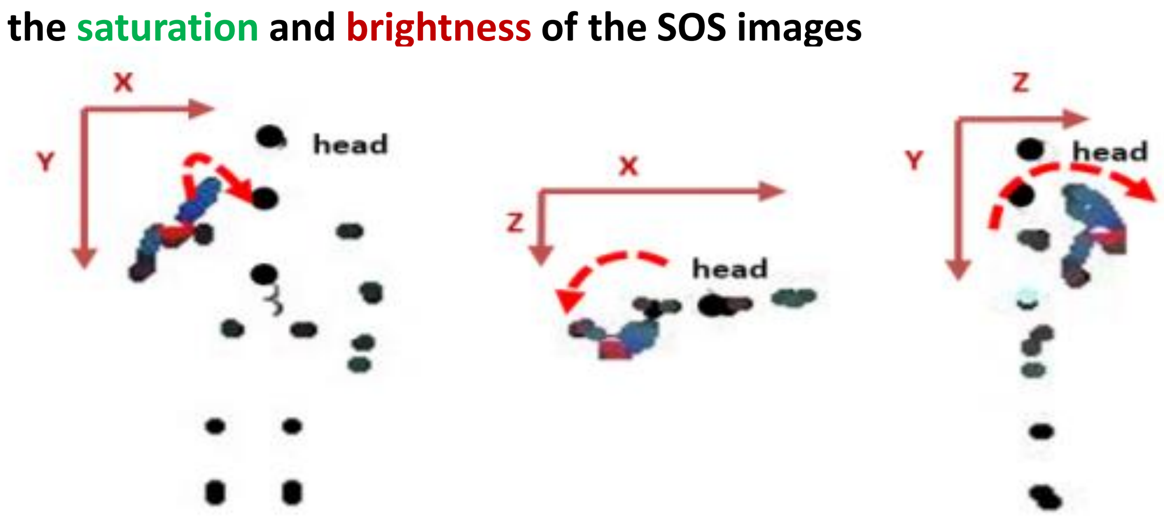
3. HAND EXTRACTION

- The hands move at different speeds showing variations in its accelerations.



4. SKELETON OPTICAL SPECTRA

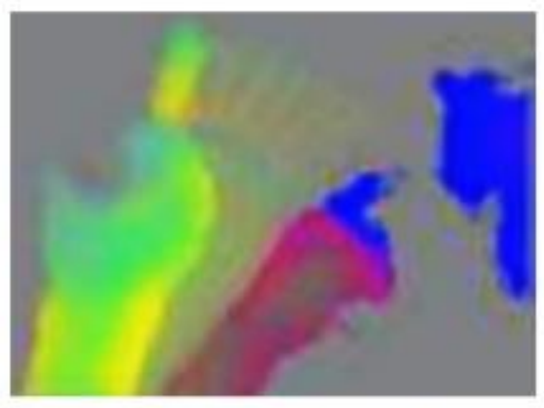
- We use the HSB color model to generate Skeleton Optical Spectra (SOS) images.
- These texture maps are capable of describing the hand movement and its location regarding the body (global movement). To further enhance the encoded spatiotemporal information, we encode the velocity of the joints into the saturation and brightness of the SOS images



(a) Global Movement and Location (skeleton)

5. RANK POOLING

- For RGB-D data, we generate dynamic images through the rank pooling method proposed by Fernando et. al. (2017) and Bilen et. al. (2016, 2017).
- The core idea is to represent a sign video through a single image that summarizes the hand's movement in the region where the sign is articulated (local movement).



(b) Local Hand Movement (Depth)

(c) Local Hand Movement (RGB)

5. EXPERIMENTS AND RESULTS

The LSA64 dataset:

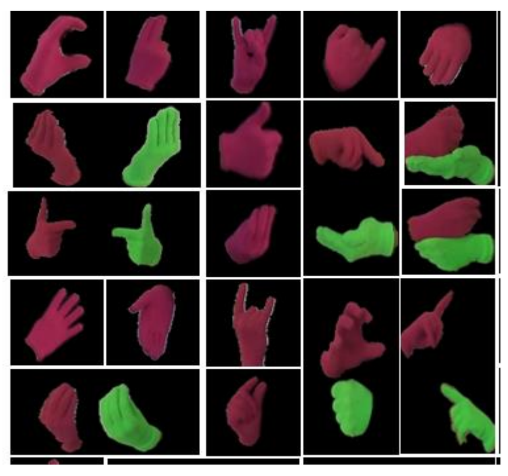
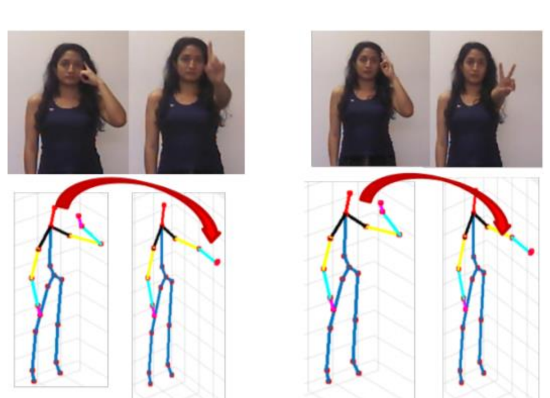


Table 1. Comparative results on the LSA64 Dataset.

Method	Accuracy (mean ± std)
ProbSOM (Ronchetti, 2018)	91.70
3DCNN (Neto et al., 2018)	93.90 ± 1.40
ALL (sequence agnostic) (Ronchetti et al., 2016b)	97.44 ± 0.59
ALL-HMM (Ronchetti et al., 2016b)	95.92 ± 0.95
Deep Network (Konstantinidis et al., 2018b)	98.09 ± 0.59
skeleton + LSTMs (Konstantinidis et al., 2018a)	99.84 ± 0.19
3S–RGBD–CNN	96.92 ± 0.56
3S–SKL–CNN	99.82 ± 0.48
Later Fusion (3S–RGBD + 3S–SKL)	99.91 ± 0.33

UFOP LIBRAS dataset:



(a) One day (b) two days

Table 2. Comparative results on the UFOP-LIBRAS Dataset.

Method	Accuracy (mean ± std)
SC-CHM (hand-crafted) (Escobedo & Camara, 2016)	63.30 ± 2.90
P-CNN (CNN + SVM) (Chéron et al., 2015)	68.14 ± 1.32
3DCNN–LSTM (Zhang et al., 2017)	74.27 ± 3.30
3S–RGBD–CNN	72.44 ± 3.35
3S–SKL–CNN	74.25 ± 3.28
Later Fusion (3S–RGBD + 3S–SKL)	75.21 ± 2.97

6. CONCLUSION

- We combined several ideas from rank pooling and skeleton optical spectra to generate texture maps to encode the location and movement of the hands.
- We proposed two multi-stream CNN models to extract spatiotemporal features of a sign.
- Experimental results showed the efficacy of the proposed method.

REFERENCES

Ronchetti, F. Reconocimiento de gestos dinámicos y su aplicación al lenguaje de señas. InXX Workshop de Investigadores en Ciencias de la Computación (WICC2018, Universidad Nacional del Nordeste)., 2018.

Cheron, G., Laptev, I., and Schmid, C. P-cnn: Pose-based cnn features for action recognition. In Proceedings of theIEEE International Conference on Computer Vision, pp.3218–3226, 2015.

Konstantinidis, D., Dimitropoulos, K., and Daras, P. Adeep learning approach for analyzing video and skeletalfeatures in sign language recognition. In2018 IEEE Inter-national Conference on Imaging Systems and Techniques(IST), pp. 1–6. IEEE, 2018a.

Escobedo, E. and Camara, G. A new approach for dy-namic gesture recognition using skeleton trajectory repre-sentation and histograms of cumulative magnitudes. InGraphics, Patterns and Images (SIBGRAPI), 2016 29thSIBGRAPI Conference on, pp. 209–216. IEEE, 2016.