Topic Detection in Continuous Sign Language Videos


Álvaro Budria


Laia Tarrés


Gerard I. Gállego


Francesc Moreno


Jordi Torres


Xavier Giro

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Institut de Robòtica i Informàtica Industrial

Intelligent Data Science and Artificial Intelligence Research Center

Barcelona Supercomputing Center Centro Nacional de Supercomputación

https://github.com/imatge-upc/sign-topic

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Accessibility, Vision, and Autonomy Workshop

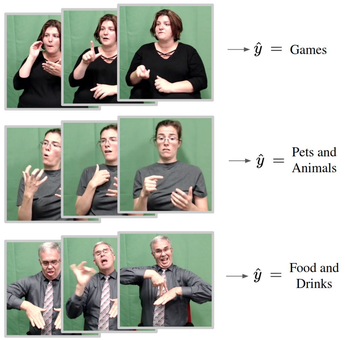
Abstract

Significant progress has been recently made on challenging tasks in automatic sign language (SL) understanding, such as sign language recognition, translation and production. However, these works have focused on datasets with relatively few samples, short recordings and limited vocabulary and signing space.

In this work, we introduce the novel task of sign language topic detection. We base our experiments on How2Sign[1], a large-scale video dataset spanning multiple semantic domains. We provide strong baselines for the task of topic detection, and present a comparison between different visual features commonly used in the domain of sign language.

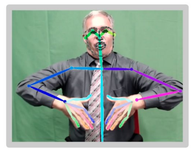
Topic Detection

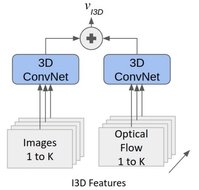
Predict a label that describes the semantic content of a signer's discourse.



Visual Features

We compare four different visual features commonly used in sign language understanding:


Cartesian Body Pose Keypoints



I3D Features

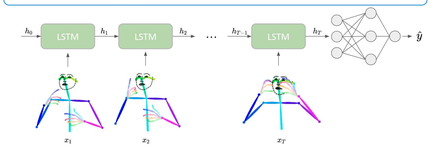

Rotational Body Pose Keypoints

Spotted Signs

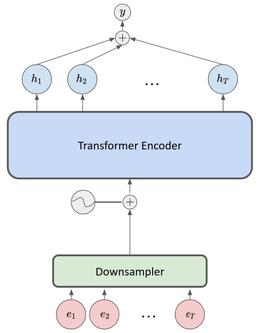
Neural Architectures

We compare three neural architectures:

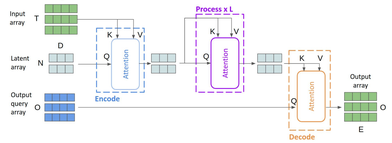
LSTM with Attention



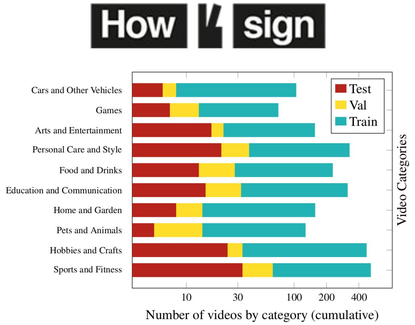
Transformer



Perceiver IO [3]

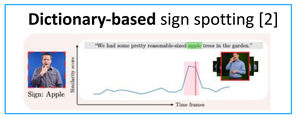


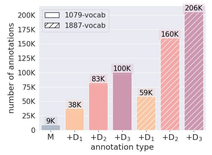
Topics

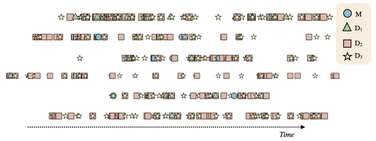


Spotted Signs [2]

We used the automatic sparse annotations available for How2Sign.





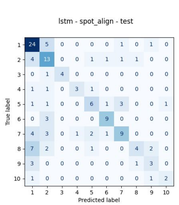


Experimental Results

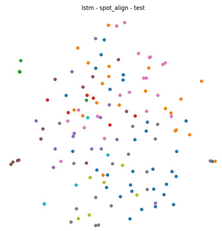
Test accuracy for each combination of features and architecture.

	LSTM	Transformer	PerceiverIO
Cartesian (OP)	30.35 ± 3.01	34.02 ± 0.33	30.34 ± 2.58
Cartesian (MP)	29.43 ± 1.17	33.10 ± 2.58	33.56 ± 1.30
Angular (OP)	31.95 ± 1.05	29.66 ± 0.12	31.49 ± 2.34
Angular (MP)	32.64 ± 0.32	34.71 ± 1.42	30.80 ± 1.97
I3D	45.75 ± 1.59	46.26 ± 1.30	48.27 ± 0.33
Spotted signs	**58.03 ± 1.40**	**53.33 ± 2.18**	**52.88 ± 0.32**
Transcriptions *(upper bound)*	70.35 ± 4.50	75.38 ± 0.75	75.90 ± 2.48

Visualizations of the spotted signs + LSTM results:



Confusion matrix



t-SNE

[1] Duarte, Amanda, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i-Nieto. "How2Sign: a large-scale multimodal dataset for continuous American sign language."CVPR 2021

[2] Duarte, Amanda, Samuel Albanie, Xavier Giró-i-Nieto, and Gül Varol. "Sign Language Video Retrieval with Free-Form Textual Queries." CVPR 2022.

[3] Jaegle, Andrew, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula et al. "Perceiver io: A general architecture for structured inputs & outputs." ICLR 2022.

Implemented in: FAIRSEQ PyTorch

Acknowledgements: Generalitat de Catalunya Departament d'Empresa i Coneixement; Ministerio de Ciencia e Innovación; Agencia Estatal de Investigación