American Sign Language Recognition

Mar Galiana Fernández

Supervisor: Abhir Bhalerao

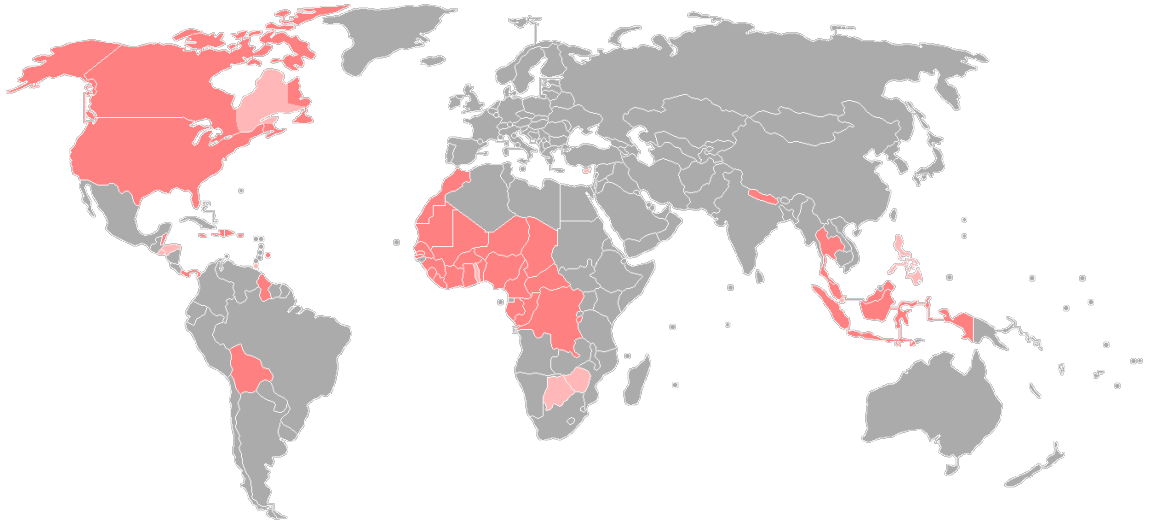
Warwick University

MSc Computer Science

TABLE OF CONTENTS

INTRODUCTION

- More than 300 different Sign Languages[1]
  - American Sign Language (ASL)



Areas where ASL is the national sign language [2]

Areas where ASL is in significant use [2]

[1] According to the World Federation of the Deaf.

[2] According to the Max Planck Institute for Evolutionary Anthropology Glottolog initiative.

Computer

PARAMETERS	THINK	DISAPPOINTED
Handshape	Closed fist with index finger extended	
Palm Orientation	Facing signer's body	
Location	Tip of finger in contact with forehead	Tip of finger in contact with chin
Movement	Unidirectional single contacting movement	

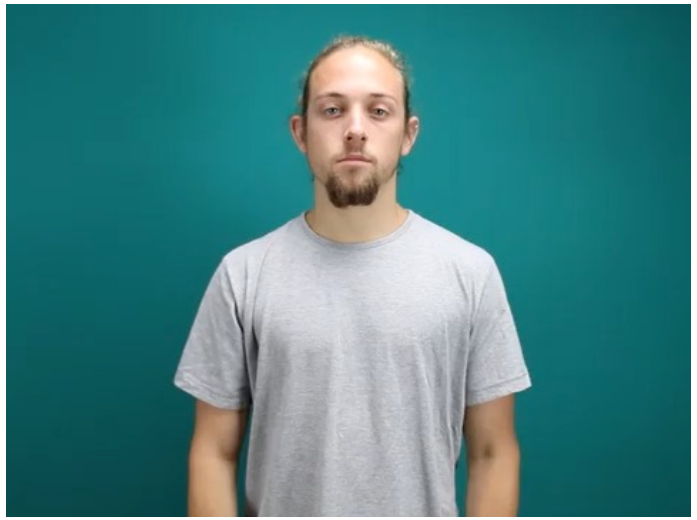
AMERICAN SIGN LANGUAGE

Parameters involved:

- Handshape
- Palm orientation
- Location
- Movement (static or dynamic)
- Nonmanual features

According to the non-manual realization of agreement in American sign language dissertation by the Boston University (http://louis-xiv.bu.edu/pub/asl/disserts/Bahan96.pdf)

Help



PARAMETERS	THINK	DISAPPOINTED
Handshape	Closed fist with index finger extended	
Palm Orientation	Facing signer's body	
Location	Tip of finger in contact with forehead	Tip of finger in contact with chin
Movement	Unidirectional single contacting movement	

AMERICAN SIGN LANGUAGE

Parameters involved:

- Handshape
- Palm orientation
- Location
- Movement (static or dynamic)
- Nonmanual features

According to the non-manual realization of agreement in American sign language dissertation by the Boston University (http://louis-xiv.bu.edu/pub/asl/disserts/Bahan96.pdf)

Not yet

PARAMETERS	THINK	DISAPPOINTED
Handshape	Closed fist with index finger extended	
Palm Orientation	Facing signer's body	
Location	Tip of finger in contact with forehead	Tip of finger in contact with chin
Movement	Unidirectional single contacting movement	

AMERICAN SIGN LANGUAGE

Parameters involved:

- Handshape
- Palm orientation
- Location
- Movement (static or dynamic)
- Nonmanual features

According to the non-manual realization of agreement in American sign language dissertation by the Boston University (http://louis-xiv.bu.edu/pub/asl/disserts/Bahan96.pdf)

MOTIVATION

1. 5% of the world's population requires rehabilitation to address their hearing loss.
   - 93% of them are adults.
   - 7% of them are children.

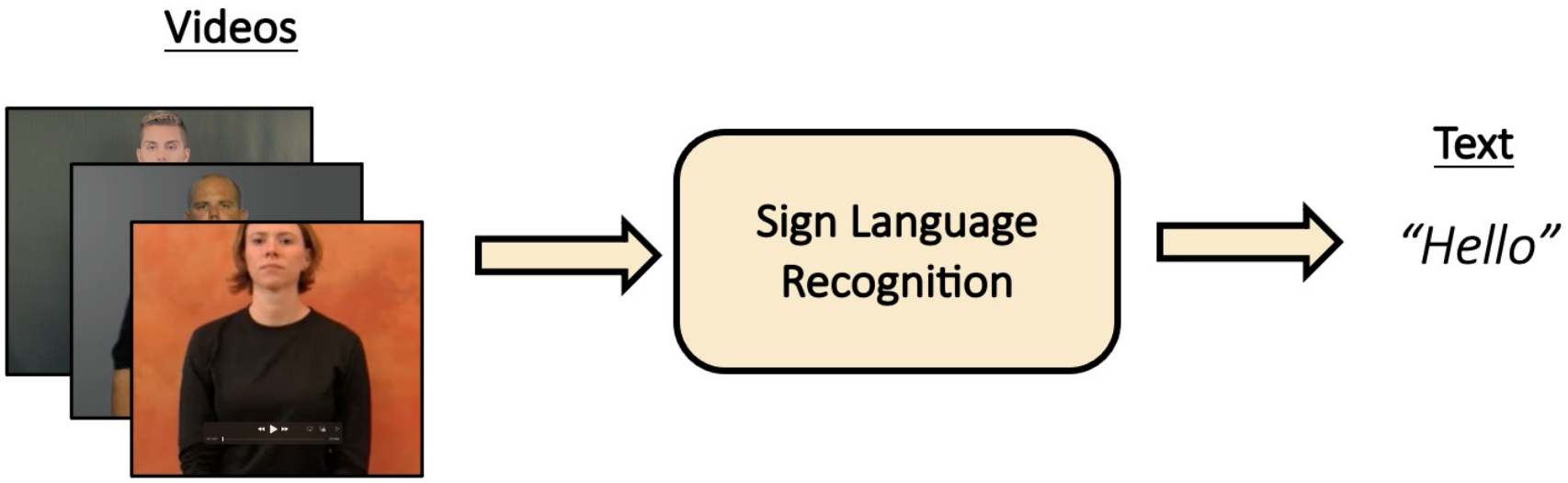
2. Reason of the Sign language chosen:
   - ASL is the most used around the world.
   - Data available.

3. Undergraduate dissertation project

According to the World Health Organization.

PROPOSAL

Implement an American Sign Language Recognition model:



SYSTEM OVERVIEW

Data Acquisition

- Cameras
  - Videos
  - Images
- Sensors
  - Arm sensors
  - Leap motion

Types of sign

- Dynamic
- Static

Recognition model

- Convolutional Neural Network
- Support Vector Machine (SVM)
- Recurrent Neural Network
- ...

RELATED WORK (1/3)

Paper	Data Acquisition	Gestures	Technique	Rate
Oz and leu [1]	Gloves	Dynamic	NN	**95%**
Sun et al [2]	Kinect	Dynamic	Latent SVM	86%
Sun et al. [3]	Kinect	Dynamic	Adaboost	86.8%
Jangyodsuk et al. [4]	Camera	Both	DTW	**93.38%**
	Kinect	Both	DTW	92.54%
Wu et al. [5]	Arm sensors	Dynamic	Decision tree	81.88%
			SVM	**99.09%**
			NN	**98.56%**
Usachokcharoen et al. [6]	Kinect	Dynamic	SVM	**95%**
Savur and Sahin [7]	Arm band	Both	SVM	82.3%
Sun et al. [8]	Kinect	Both	Latent SVM	86%
Kumar et al. [9, 10]	Camera	Static	SVM	**93%**
		Dynamic	SVM	**100%**
Savur and Sahin [11]	Arm band	Dynamic	SVM	60.85%
Al-Hammadi et al. [14]	Camera	Dynamic	3DCNN + MLP	84.38%
			LSTM	82.19%
Hassan et al. [15]	Camera	Dynamic	SlowFast Network	79.34%

Summary of the Sign Language Recognition systems implemented until the 2019 by Wadhawan and Kumar [13]

RELATED WORK (2/3)

Kumar et al. [9, 10] (2016)

Sign Language: American

Classes: Alphabet (24 letters)

- j and z dynamic

Techniques:

- Viola-Jones face detection
- SVM

Results:

- Dynamic labels for testing: (j, z, no, goodbye)
- Accuracy: 100% for dynamic signs

Jangyodsuk et al. [4] (2014)

Sign Language: American

Classes: 3 signers, each 1,113 signs.

Techniques:

- Dynamic Time Warping (DTW) to compare the hand trajectory
- Histogram of Oriented Gradient (HoG) improved accuracy up to 8%

Results:

- Best accuracy: 93.38%

Al-Hammadi et al. [14] (2020)

Sign Language: Saudi

Classes: 40 dynamic words

Techniques:

- 3D Convolutional Neural Network (3DCNN)
- Multilayer Perceptron (MLP)
- Stacked autoencoder.

Results:

- Accuracy (3DCNN and MLP): 84.38%

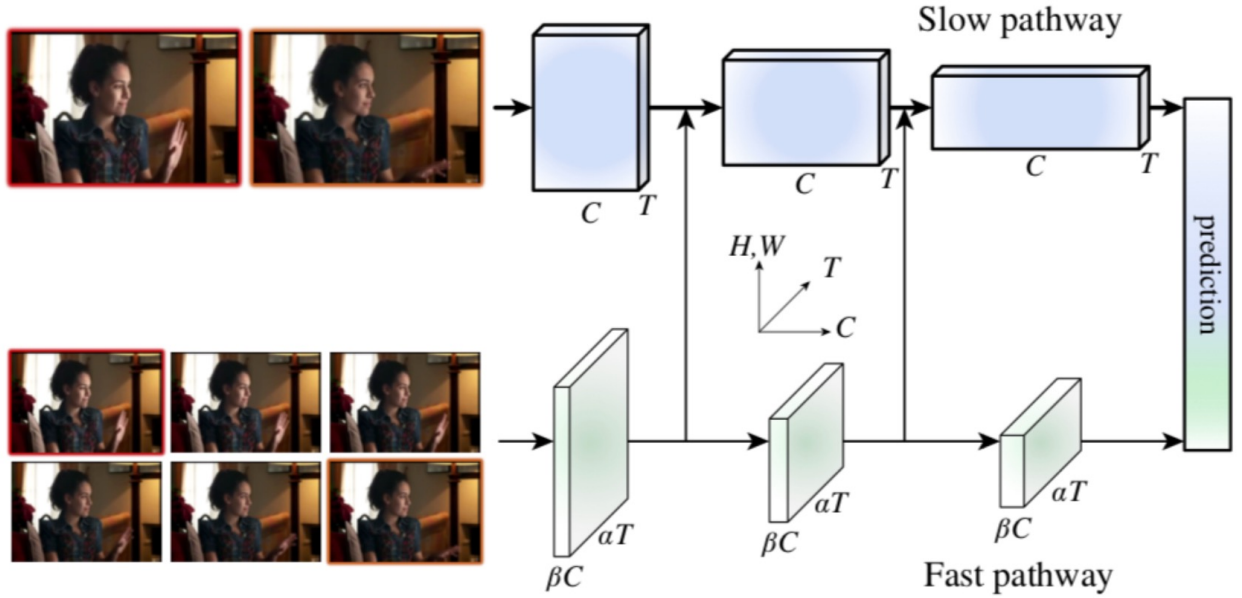
RELATED WORK (3/3)

Enhanced Dynamic Sign Language Recognition using SlowFast Networks By Hassan et al. [15]

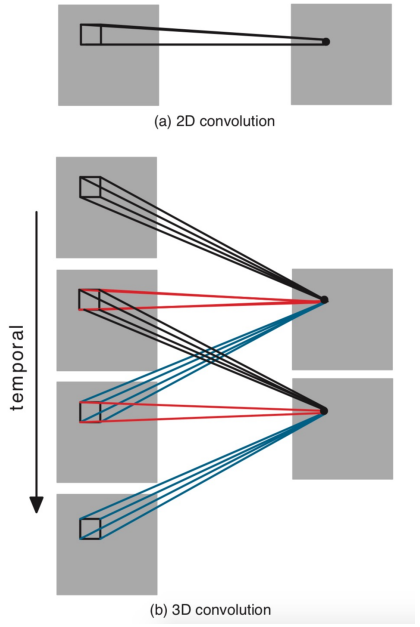
- **Dataset:** Word-Level American Sign Language (WLASL)
- **Number of classes:** 300 labels (30 fps)
- **Model:** Pre-trained model SLOWFAST_8×8_R50 by the PySlow Fast Github
- **Limitations:**
  - Hardware limitations
  - Time consuming (24 hours training)
- **Results:** Improvement of 23.2% over the previous state-of-the-art of the

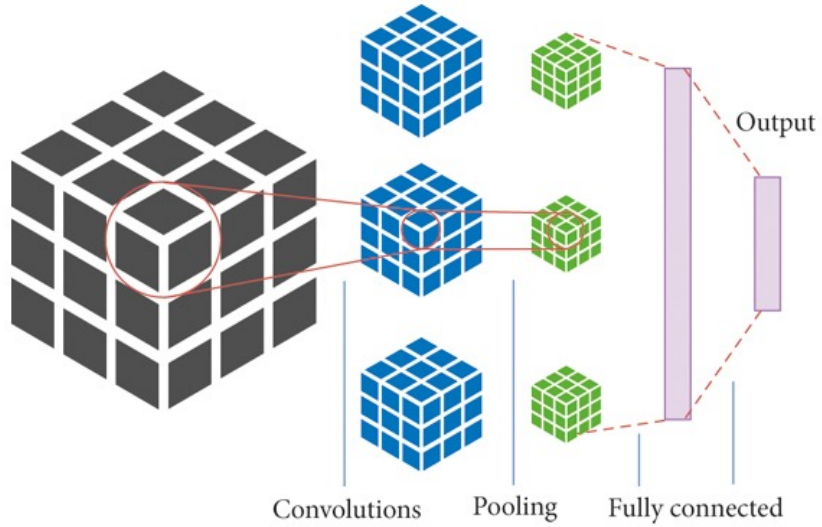
	Top 1	**Top 5**
Pose-GRU	33.68	64.37
Pose-TGCN	38.32	67.51
VGG-GRU	19.31	46.56
I3D	56.14	79.94
SlowFast (proposed)	**79.34**	**90.31**

SLOWFAST NETWORK



3D CONVOLUTIONAL NEURAL NETWORK





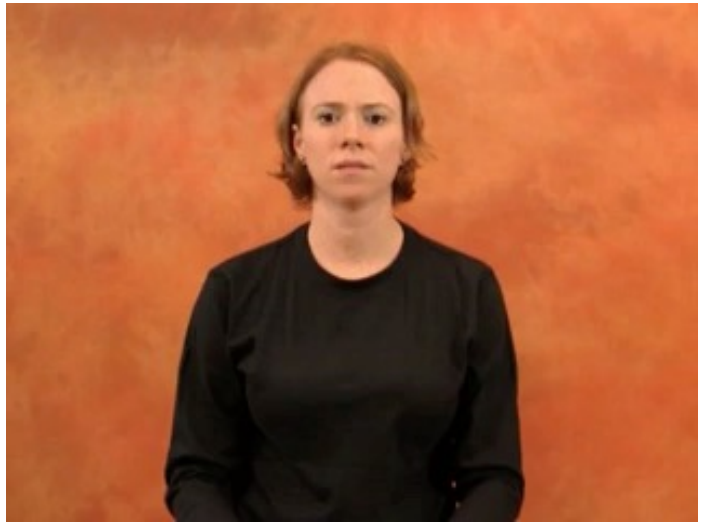
WORK DONE (1/5)

Background Research

Dataset selection

Experiments

WORK DONE (2/5): DATASET SELECTION



Word-Level American Sign Language (WLASL)

Dataset	#Classes	#Videos	Mean	#Signers
WLASL100	100	2,038	20.4	97
WLASL300	300	5,117	17.1	109
WLASL1000	1,000	13,168	13.2	116
WLASL2000	2,000	21,083	10.5	119

WORK DONE (3/5): EXPERIMENTS

Model Name	Top-1 Accuracy	Top-5 Accuracy	Input Shape
MoViNet-A0-Base	72.28	90.92	50 x 172 x 172
MoViNet-A1-Base	76.69	93.40	50 x 172 x 172
MoViNet-A2-Base	78.62	94.17	50 x 224 x 224
MoViNet-A3-Base	81.79	95.67	120 x 256 x 256
MoViNet-A4-Base	83.48	96.16	80 x 290 x 290
MoViNet-A5-Base	84.27	96.39	120 x 320 x 320

Model: MoViNet-A0-Base

Classes:

- 10 words (drink, trade, before, bowling, computer, cool, go, thin, help and tall)
- 8 videos per class

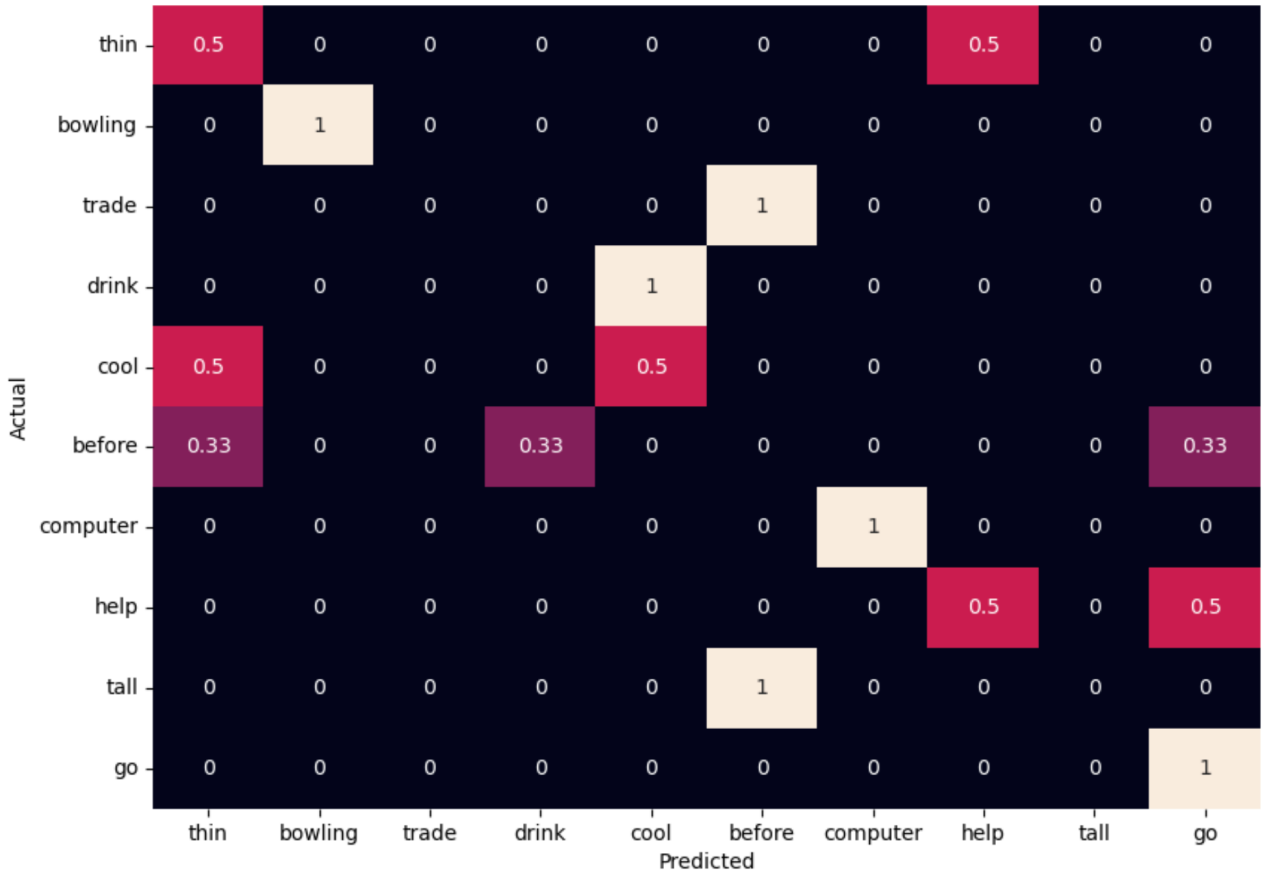
Parameters:

- Batch size: 8
- Epochs: 10
- Frames per video: 20

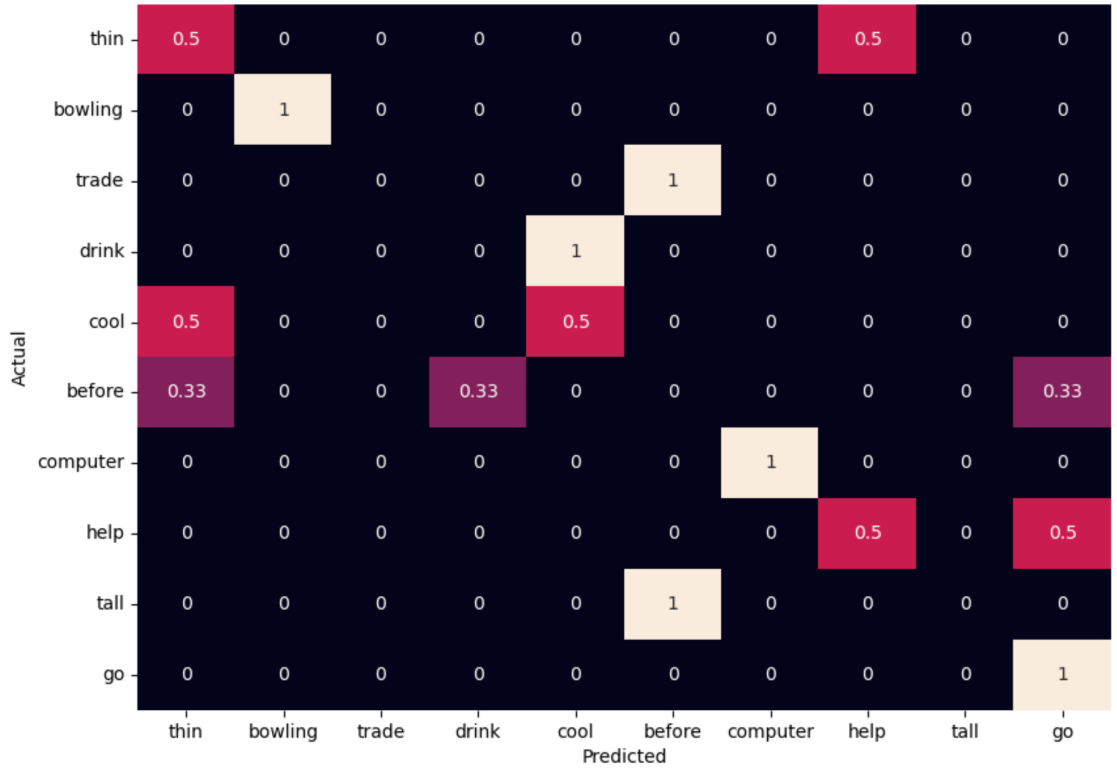
Model is obtained from the experiments done in the MoViNets: Mobile Video Networks for Efficient Video Recognition paper by Kondratyuk et al. [16]

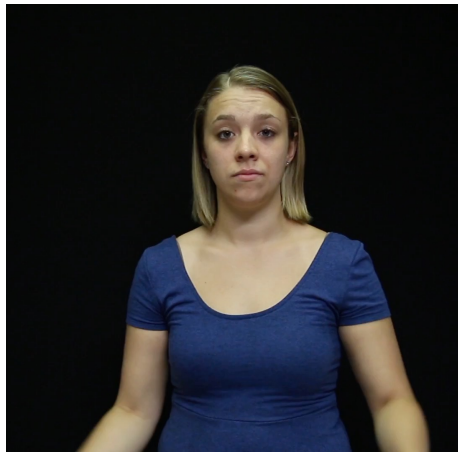
WORK DONE (4/5): EXPERIMENTS

Experiment	Frames position	Accuracy
1	Beginning	40%
2	Middle	40%
3	End	20%
4	Random start	40%

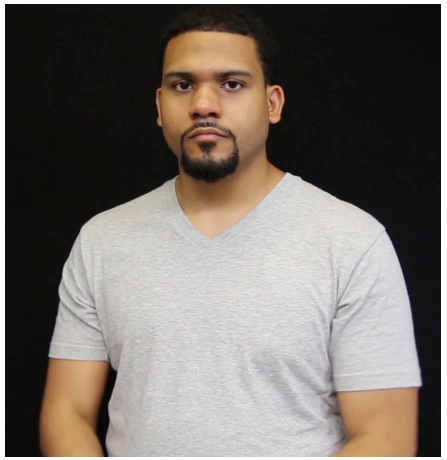


WORK DONE (5/5): EXPERIMENTS

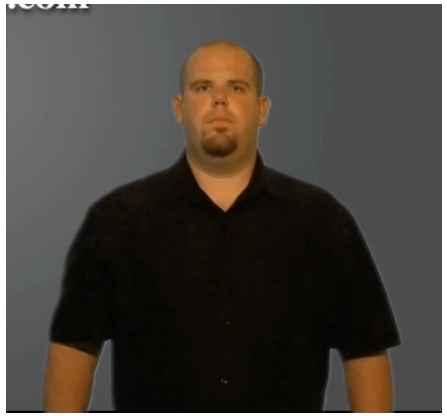




Trade



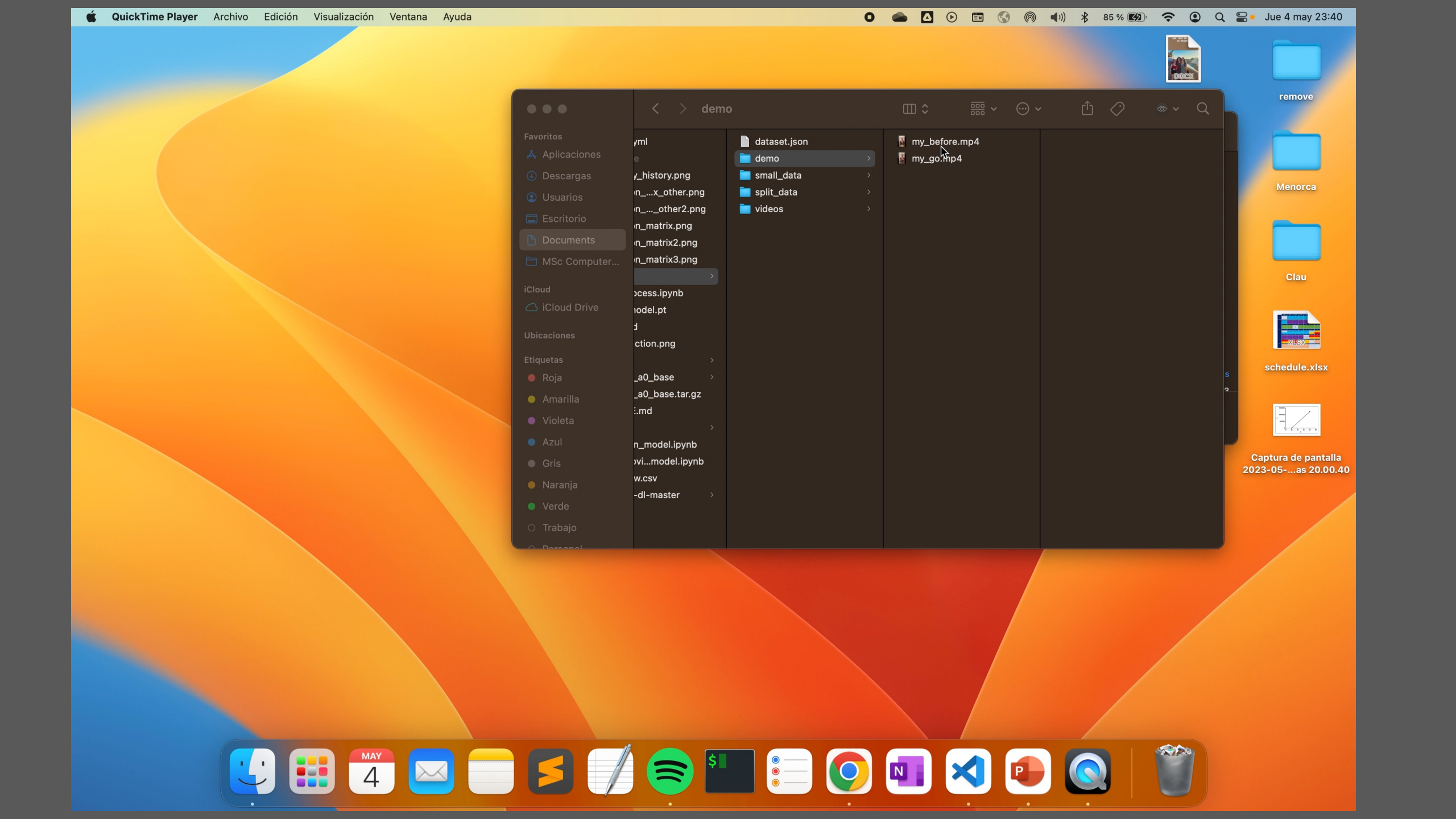
Tall

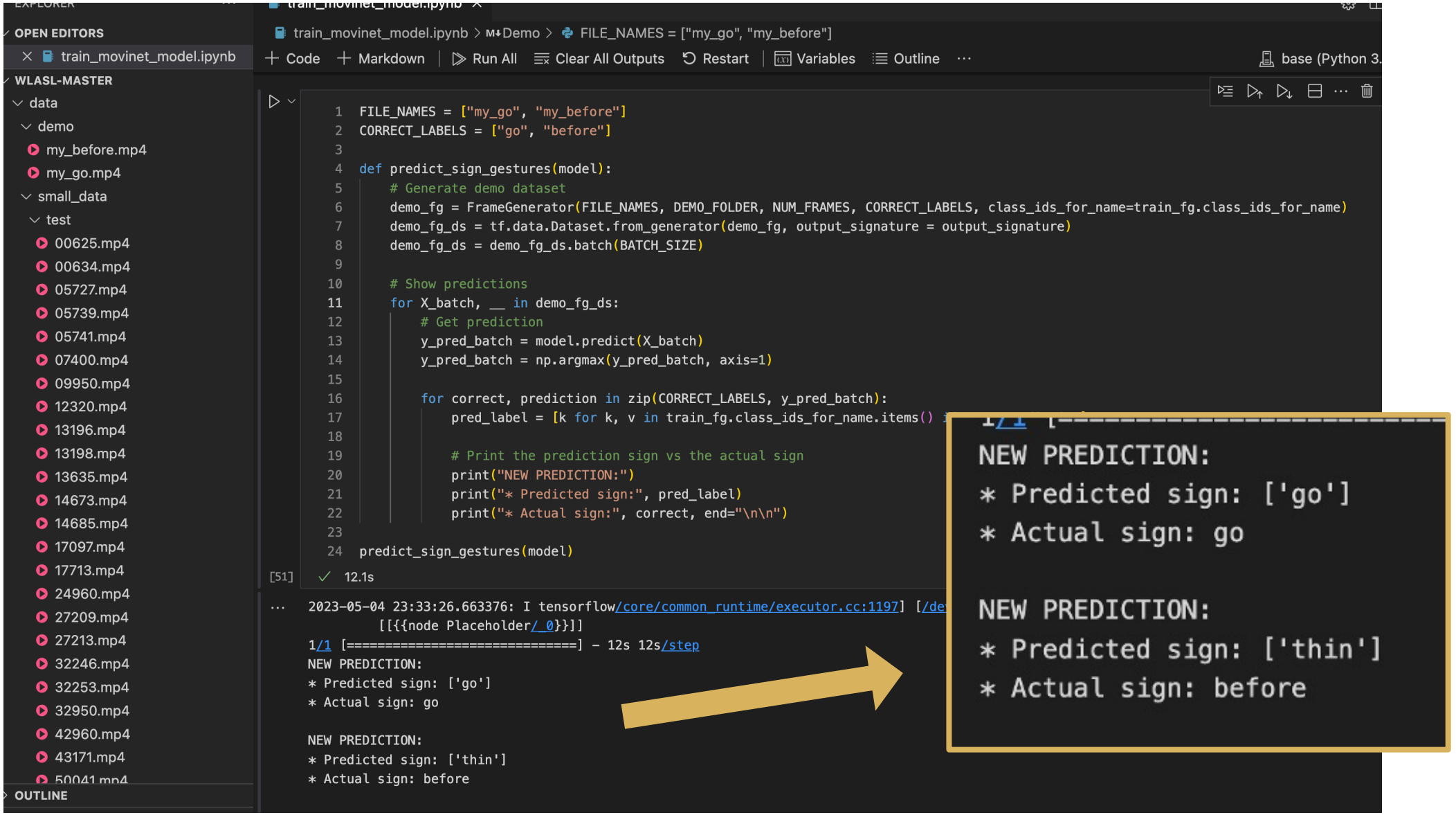


Before

Model is obtained from the experiments done in the MoViNets: Mobile Video Networks for Efficient Video Recognition paper by Kondratyuk et al. [16]

VIDEO


QuickTime Player
Archivo
Edición
Visualización
Ventana
Ayuda
Jue 4 may 23:40
demo
Favoritos
Aplicaciones
Descargas
Usuarios
Escritorio
Documents
MSc Computer...
iCloud
iCloud Drive
Ubicaciones
Etiquetas
Roja
Amarilla
Violeta
Azul
Gris
Naranja
Verde
Trabajo
dataset.json
demo
small_data
split_data
videos
my_before.mp4
my_go.mp4
remove
Menorca
Clau
schedule.xlsx
Captura de pantalla 2023-05-...as 20.00.40


OPEN EDITORS
train_movinet_model.ipynb
WLASL-MASTER
data
demo
my_before.mp4
my_go.mp4
small_data
test
00625.mp4
00634.mp4
05727.mp4
05739.mp4
05741.mp4
07400.mp4
09950.mp4
12320.mp4
13196.mp4
13198.mp4
13635.mp4
14673.mp4
14685.mp4
17097.mp4
17713.mp4
24960.mp4
27209.mp4
27213.mp4
32246.mp4
32253.mp4
32950.mp4
42960.mp4
43171.mp4
50041.mp4
OUTLINE
train_movinet_model.ipynb > Demo > FILE_NAMES = ["my_go", "my_before"]
Code
Markdown
Run All
Clear All Outputs
Restart
Variables
Outline
base (Python 3.
FILE_NAMES = ["my_go", "my_before"]
CORRECT_LABELS = ["go", "before"]
def predict_sign_gestures(model):
# Generate demo dataset
demo_fg = FrameGenerator(FILE_NAMES, DEMO_FOLDER, NUM_FRAMES, CORRECT_LABELS, class_ids_for_name=train_fg.class_ids_for_name)
demo_fg_ds = tf.data.Dataset.from_generator(demo_fg, output_signature = output_signature)
demo_fg_ds = demo_fg_ds.batch(BATCH_SIZE)
# Show predictions
for X_batch, __ in demo_fg_ds:
# Get prediction
y_pred_batch = model.predict(X_batch)
y_pred_batch = np.argmax(y_pred_batch, axis=1)
for correct, prediction in zip(CORRECT_LABELS, y_pred_batch):
pred_label = [k for k, v in train_fg.class_ids_for_name.items()
# Print the prediction sign vs the actual sign
print("NEW PREDICTION:")
print("* Predicted sign:", pred_label)
print("* Actual sign:", correct, end="\n\n")
predict_sign_gestures(model)
[51]
12.1s
2023-05-04 23:33:26.663376: I tensorflow/core/common_runtime/executor.cc:1197] [/de
[[{{node Placeholder/_0}}]]
1/1 [==============================] - 12s 12s/step
NEW PREDICTION:
* Predicted sign: ['go']
* Actual sign: go
NEW PREDICTION:
* Predicted sign: ['thin']
* Actual sign: before
NEW PREDICTION:
* Predicted sign: ['go']
* Actual sign: go
NEW PREDICTION:
* Predicted sign: ['thin']
* Actual sign: before

PLAN OF WORK

- Implement SlowFast Network model
- Data augmentation
- Use of other datasets: MS-ASL

MILESTONS	
Dissertation Specification - Research Proposal	2nd March 2023
Presentation	4th May 2023
Last exam	10th June 2023
Interim Report	13th July 2023
Dissertation Report	7th September 2023

Conclusion

Thank you for your attention

Are there any questions?

REFERENCES (1/2)

[1] Cemil Oz and Ming C. Leu. Linguistic properties based on american sign language isolated word recognition with artificial neural networks using a sensory glove and motion tracker. Neurocomputing, 70(16):2891–2901, 2007. Neural Network Applications in Electrical Engineering Selected papers from the 3rd International Work-Conference on Artificial Neural Networks (IWANN 2005).

[2] Chao Sun, Tianzhu Zhang, Bingkun Bao, and Changsheng Xu. Latent support vector machine for sign language recognition with kinect. 2013 IEEE International Conference on Image Processing, pages 4190–4194, 2013.

[3] Chao Sun, Tianzhu Zhang, Bing-Kun Bao, Changsheng Xu, and Tao Mei. Discriminative exemplar coding for sign language recognition with kinect. IEEE Transactions on Cybernetics, 43(5):1418–1428, 2013.

[4] Pat Jangyodsuk, Christopher Conly, and Vassilis Athitsos. Sign lan- guage recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features. In Proceedings of the 7th International Conference on PErvasive Technologies Related to As- sistive Environments, PETRA ’14, New York, NY, USA, 2014. Associ- ation for Computing Machinery.

[5] Jian Wu, Zhongjun Tian, Lu Sun, Leonardo Estevez, and Roozbeh Ja- fari. Real-time american sign language recognition using wrist-worn mo- tion and surface emg sensors. In 2015 IEEE 12th International Confer- ence on Wearable and Implantable Body Sensor Networks (BSN), pages 1–6, 2015.

[6] Panupon Usachokcharoen, Yoshikazu Washizawa, and Kitsuchart Pasupa. Sign language recognition with microsoft kinect’s depth and colour sensors. 2015 IEEE International Conference on Signal and Im- age Processing Applications (ICSIPA), pages 186–190, 2015.

[7] Celal Savur and Ferat Sahin. Real-time american sign language re- cognition system using surface emg signal. 2015 IEEE 14th Inter- national Conference on Machine Learning and Applications (ICMLA), pages 497–502, 2015.

[8] Chao Sun, Tianzhu Zhang, and Changsheng Xu. Latent support vec- tor machine modeling for sign language recognition with kinect. ACM Trans. Intell. Syst. Technol., 6(2), mar 2015.

[9] Anup Kumar, Karun Thankachan, and Mevin M. Dominic. Sign lan- guage recognition. 2016 3rd International Conference on Recent Ad- vances in Information Technology (RAIT), pages 422–428, 2016.

REFERENCES (2/2)

[10] D. Anil Kumar, Polurie Venkata Vijay Kishore, A. S. Chandrasekhara Sastry, and P. Reddy Gurunatha Swamy. Selfie continuous sign language recognition using neural network. 2016 IEEE Annual India Conference (INDICON), pages 1–6, 2016.

[11] Celal Savur and Ferat Sahin. American sign language recognition system by using surface emg signal. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 002872–002877, 2016.

[12] Duaa AlQattan and Francisco Sepulveda. Towards sign language recog- nition using eeg-based motor imagery brain computer interface. 2017 5th International Winter Conference on Brain-Computer Interface (BCI), pages 5–8, 2017.

[13] Ankita Wadhawan and Parteek Kumar. Sign language recognition sys- tems: A decade systematic literature review. Archives of Computational Methods in Engineering, 28:785 – 813, 2019.

[14] Muneer Al-Hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsulaiman, Mohamed A. Bencherif, and Mohamed Amine Mekhtiche. Hand gesture recognition for sign language using 3dcnn. IEEE Access, 8:79491–79509, 2020.

[15] A. Hassan, A. Elgabry and E. Hemayed, "Enhanced Dynamic Sign Language Recognition using SlowFast Networks," 2021 17th International Computer Engineering Conference (ICENCO), Cairo, Egypt, 2021, pp. 124-128, doi: 10.1109/ICENCO49852.2021.9698904.

[16] Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., & Gong, B. (2021). Movinets: Mobile video networks for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16020-16030).

QUESTIONS SUPPORT

DATASET INFORMATION

VIDEO DETAILS

- Number of videos: 1694
- Frames per second: 23 – 40
- Data velocity: 2,35 Mbit/s
- Proportions: 16:9
- Resolution: 1920 × 1080
- Extension: mp4

NUMBER OF VIDEOS AND OCCURRENCE OF THE SAME LABEL IN EACH SET

- Train
  - Amount of data: 1242 . Percentage: 73.32 %
  - Number of different types of labels: 185
  - Different types of labels occurrence: [Min: 3 , Max: 12]
- Test
  - Amount of data: 164 . Percentage: 9.68 %
  - Number of different types of labels: 123
  - Different types of labels occurrence: [Min: 1 , Max: 3]
- Validation
  - Amount of data: 288 . Percentage: 17.00 %
  - Number of different types of labels: 164
  - Different types of labels occurrence: [Min: 1 , Max: 4]

NUMBER OF SINGERS IN EACH SET

- Train: 1 to 207
- Test: 1 to 17
- Val: 1 to 49

OCCURRENCE OF SINGERS FOR THE SAME LABEL IN EACH SET

- Train: 1 to 5
- Test: 1 to 2
- Val: 1 to 2

OCCURRENCE OF SINGERS FOR THE SAME LABEL IN DIFFERENT SETS

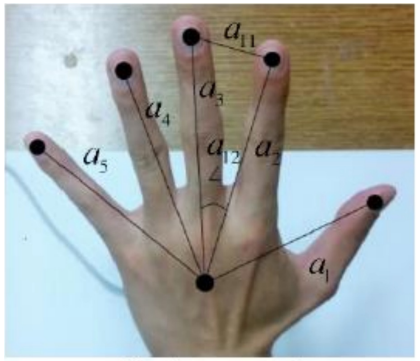
- train-test
  - Min occurrence: 0. Max occurrence: 2
  - Times occurrence: 41 . Percentage: 22.16 %
- train-val
  - Min occurrence: 0. Max occurrence: 3
  - Times occurrence: 78 . Percentage: 42.16 %
- test-val
  - Min occurrence: 0. Max occurrence: 1
  - Times occurrence: 17 . Percentage: 13.82 %
- val-test
  - Min occurrence: 0. Max occurrence: 1
  - Times occurrence: 17 . Percentage: 10.37 %

MS-ASL DATASET

Set	Classes	Subjects	Samples	Duration	Sample per class
MS-ASL100	100	189	5736	5:33	57.4
MS-ASL200	200	196	9719	9:31	48.6
MS-ASL500	500	222	17823	17:19	35.6
MS-ASL1000	1000	222	25513	24:39	25.5

DATA ACQUISITION

Leap motion controller

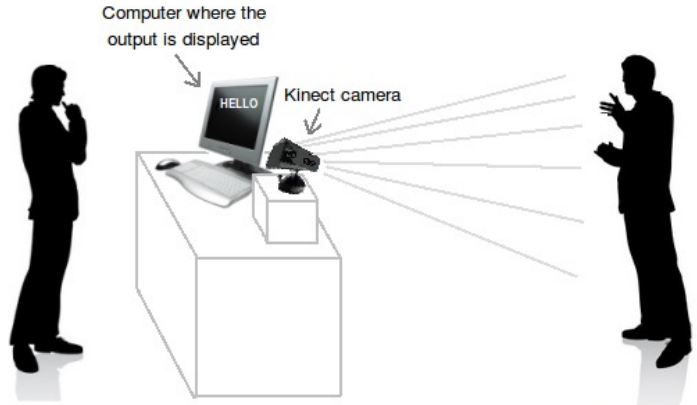


Gesture mapping

Leap motion controller hand bone

Arm Band

Kinect



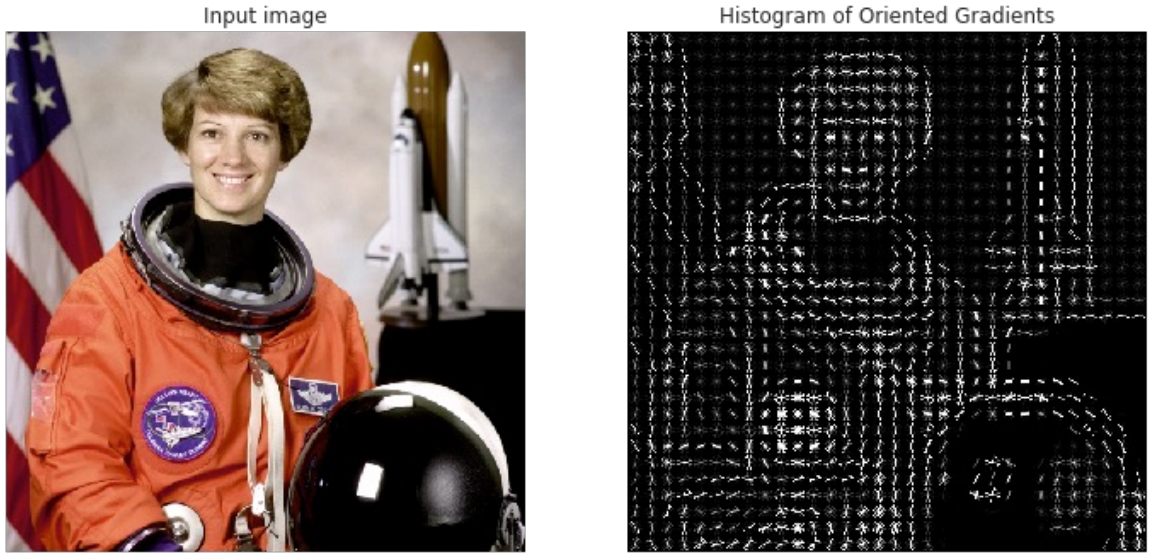
Ordinary user

Deaf user

Arm Sensor

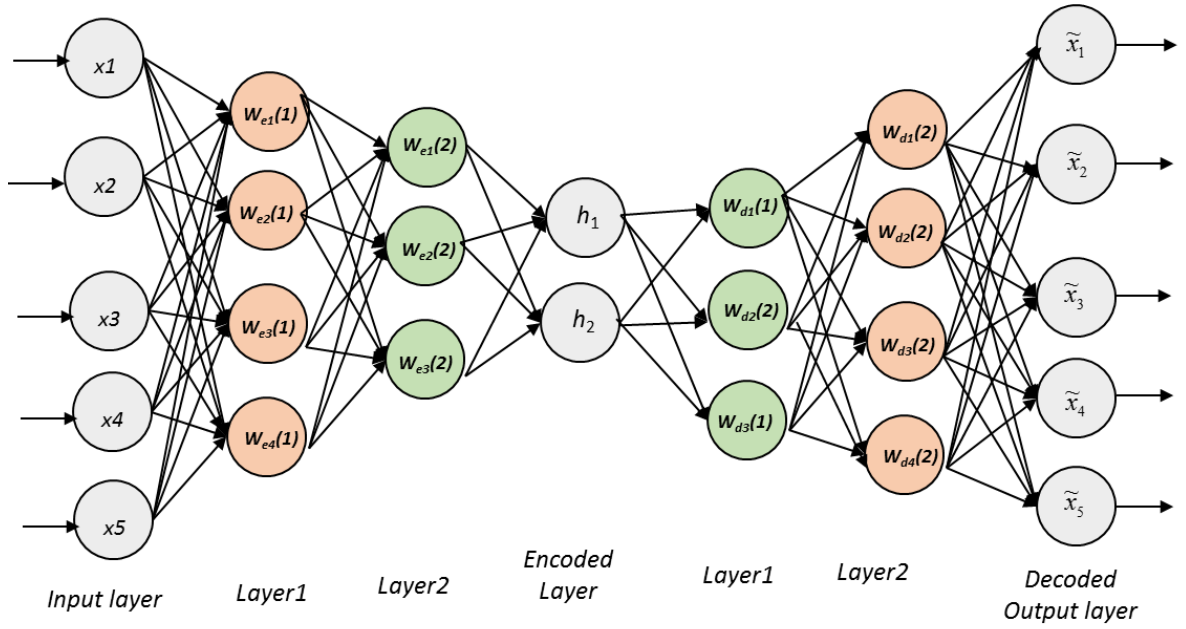
HISTOGRAM OF ORIENTED GRADIENTS (HoG)

HoG is a feature descriptor like the Canny Edge Detector, SIFT (Scale Invariant and Feature Transform) . It is used in computer vision and image processing for the purpose of object detection.



Stacked Autoencoder

Neural network of several layers of sparse autoencoders where output of each hidden layer is connected to the input of the successive hidden layer



Dynamic Time Warping

Algorithm for measuring similarity between two temporal sequences, which may vary in speed.