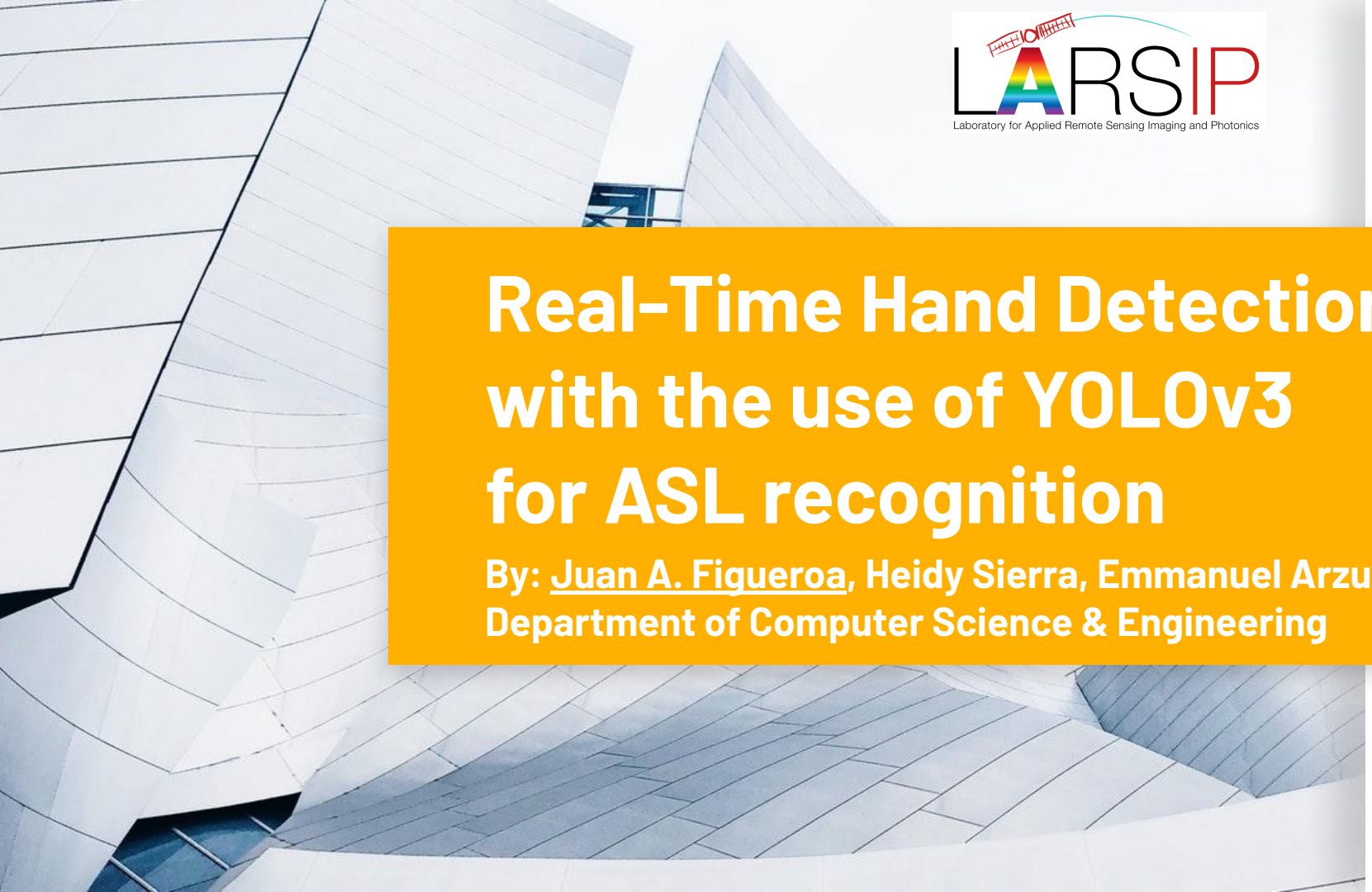


LARSIP
Laboratory for Applied Remote Sensing Imaging and Photonics

RECINTO UNIVERSITARIO DE MAYAGÜEZ
UNIVERSIDAD DE PUERTO RICO
CAAM

Real-Time Hand Detection with the use of YOLOv3 for ASL recognition

By: Juan A. Figueroa, Heidy Sierra, Emmanuel Arzuaga
Department of Computer Science & Engineering

Contents

- Objective
- Related Work
- A YOLO based ASL translator
- Test & Results
- Limitations
- Future Work & Conclusion

1. Objective

“

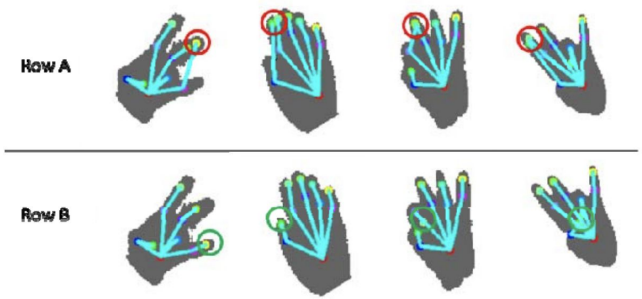
Develop a real-time ASL recognition System with the use of YOLOv3.

2. Related Work

State of the art

Kinect ASL Translator

- Kinect Sensor for ASL translation
  - Hand segmentation with depth contrast
  - Localize hand joints
  - Classify with a Random Forest Classifier
- Over 90% accuracy with 24 static ASL signs



Cao Dong, Ming C Leu, and Zhaozheng Yin. American sign language alphabet recognition using microsoft kinect. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 44–52, 2015

MIT SignAloud

Developed a specialized glove with sensors that tracked hand gestures and later sent the data through bluetooth. The computer would later receive and identify the gestures through various statistical regressions.

Lemelson-MIT.SignAloud: Gloves that Transliterate Sign Language into Text and Speech, 2017.

Cross Correlation Translator

- Extracted hand from a still frame
  - Classic image processing techniques
- Extracted features of the hand
  - Edge detectors
- Classified using cross-correlation
  - Compared with self-made hand database
- 94.23% accuracy with ASL alphabet

Anshal Joshi, Heidy Sierra, and Emmanuel Arzuaga. American sign language translation using edge detection and crosscorrelation. In 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), pages 1–6. IEEE,2017.

Convex-hull CNN Translator

- Hand extraction from still frame
  - YCbCr Skin Color Segmentation
- Hand region extraction
  - Convex-hull Jarvis's Algorithm
- Classified hand sign
  - CNN real time classification
- 98.05%accuracy on 360 test, 10 for each symbol in ASL

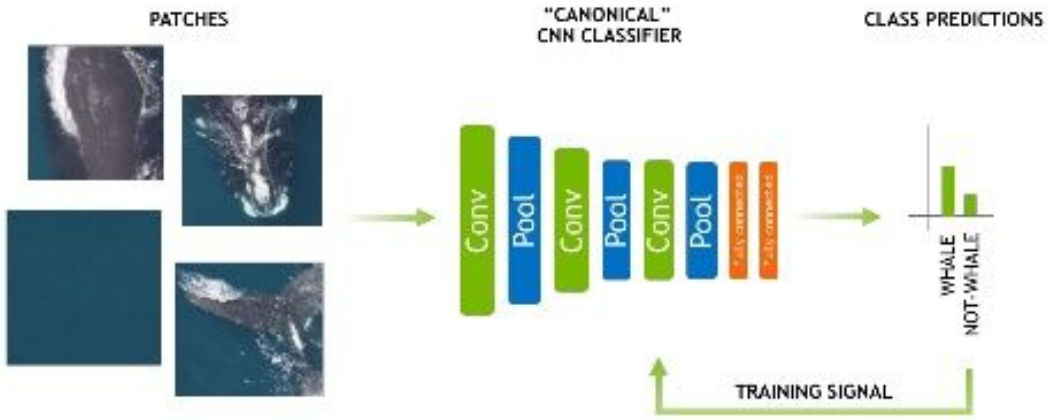
Murat Taskiran, Mehmet Killioglu, and Nihan Kahraman. A real time system for recognition of american sign language by using deep learning. In 2018 41st International Conference on Telecommunications and Signal Processing (TSP), pages 1–5.IEEE, 2018.

CNN Selection

- Deformable Parts Model (DPM)*
  - Complex pipeline
- R-CNN*
  - Region Proposal Bottleneck
- Faster R-CNN*
  - Still not fast enough

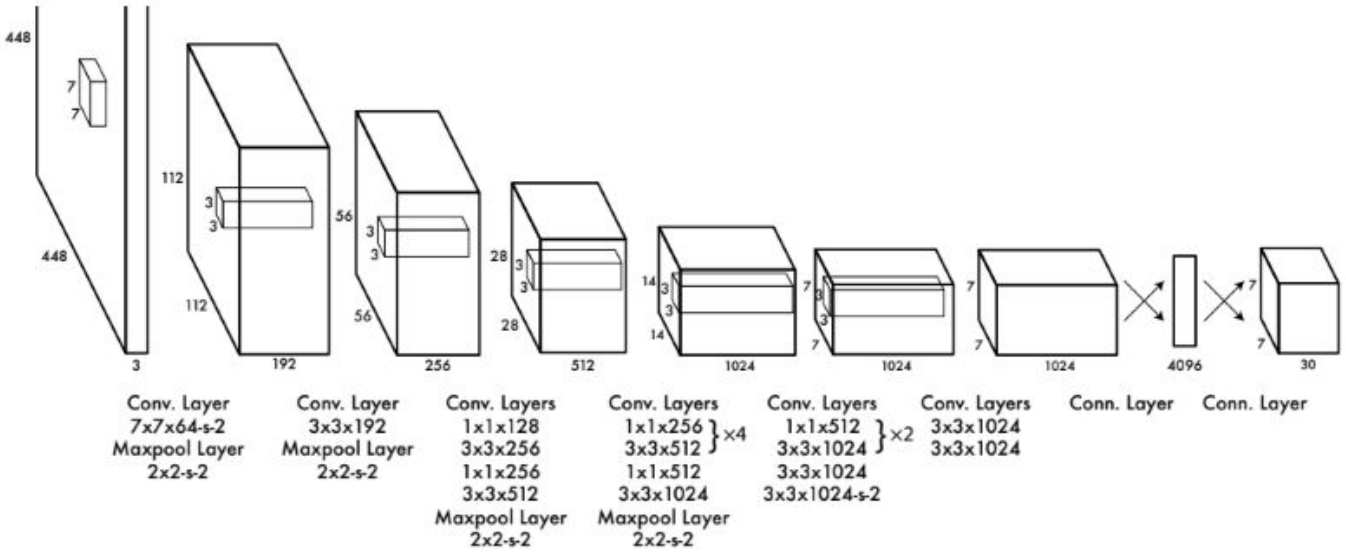
* All aforementioned networks are region or sliding window based approaches

Sliding Window Approach

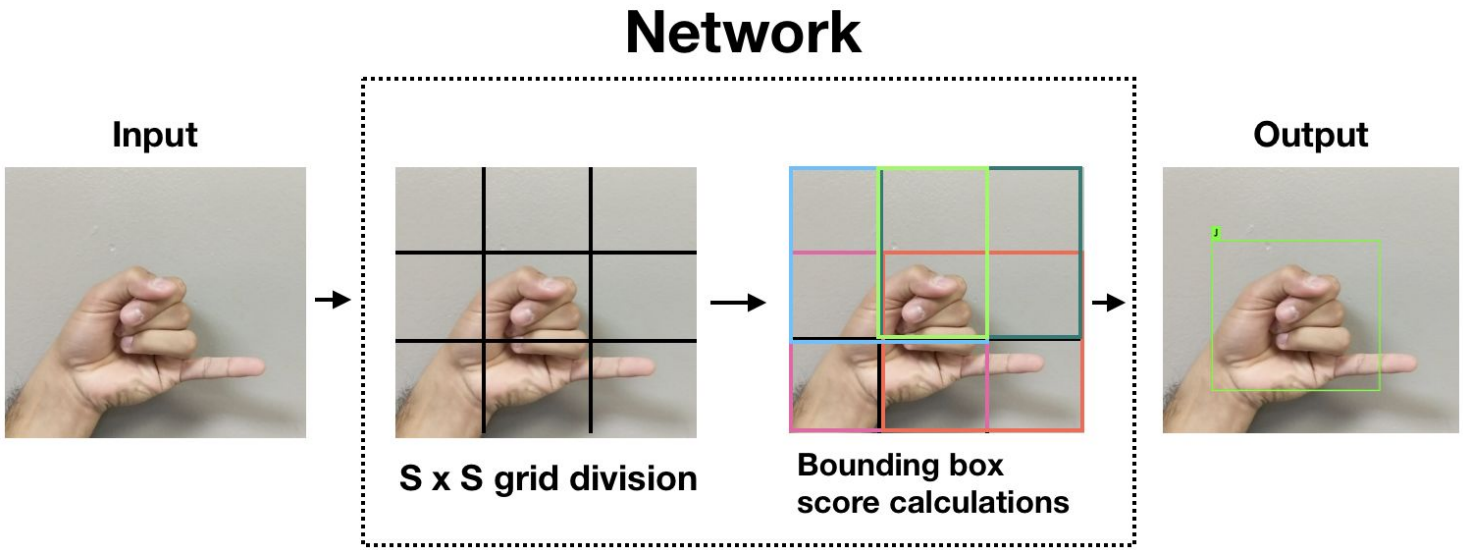


Why You Only Look Once (YOLO)?

- Generalized network
- Single network predicts and classifies
- Detects and classifies objects simultaneously
- Looks at the entire image at the time of training
- Fast detection speeds



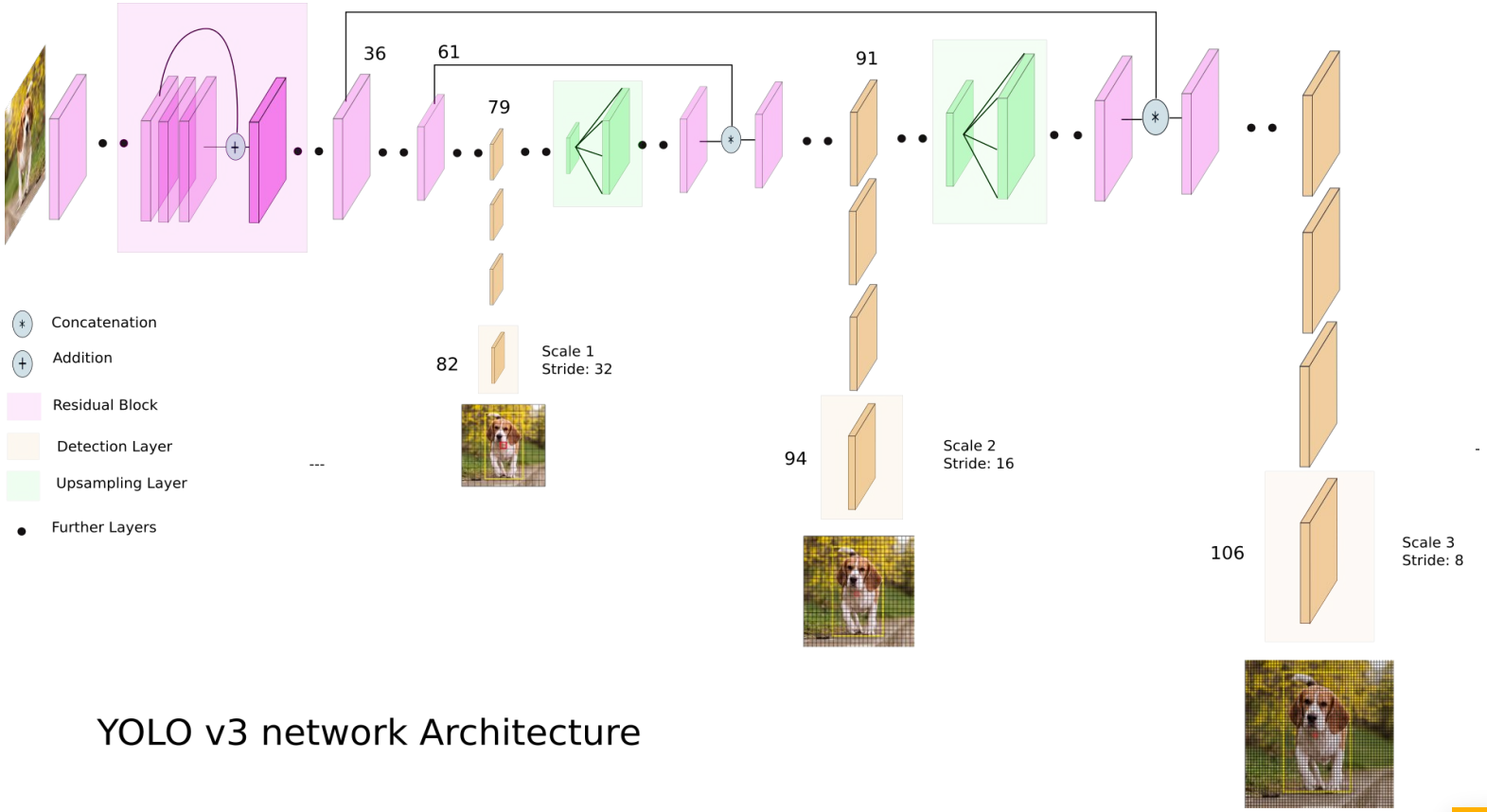
YOLO Architecture



YOLO prediction and classification

What's new in YOLOv3

- Multi-Class labeling
  - Better small Object detection
  - Generalization improvement
- Improved Overall Accuracy



YOLO v3 network Architecture

2. A YOLO based ASL translator

Experimental Platform & The System

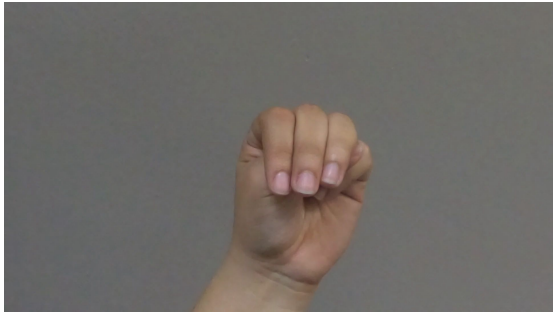
Training Dataset Description

- 26 videos, one for each ASL letter
  - Each video at least 10 seconds long
  - GoPro Hero 4 at 30 fps, 1080p resolution
- 8,900 annotated images, ~350 per letter
- 90% of dataset utilized in training

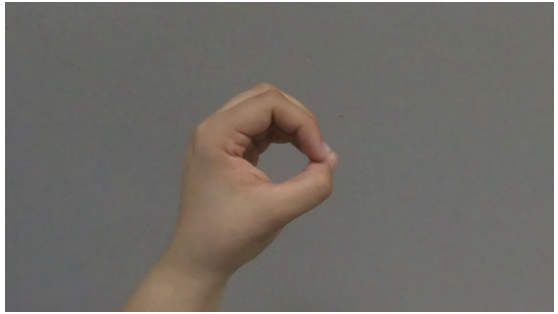
Test Dataset Description

- 130 videos, 10,521 still frames
  - 5 different individuals*
  - All 26 ASL characters
- Taken with Pixel 2 XL
  - 30 fps, 1080p resolution

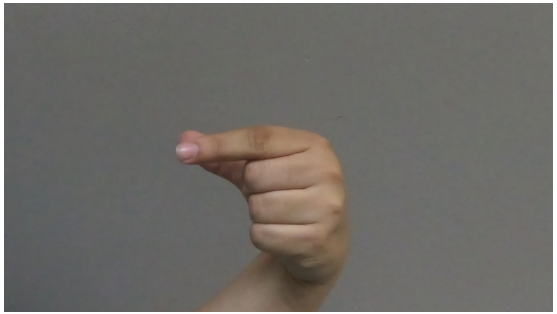
* None of the 5 individuals knew ASL before the tests.



ASL Symbol for ‘M’



ASL Symbol for ‘O’



ASL Symbol for ‘G’

ASL Dataset Examples

Experimental Setup

- PC
  - CPU - AMD® A10-7850k Radeon R7
  - GPU - GeForce GTX 1080 Ti, 11G
  - SSD - WD 1 Terabyte SSD
  - OS - Ubuntu 16.04
- Dependencies
  - Darknet
  - Python
  - CUDA

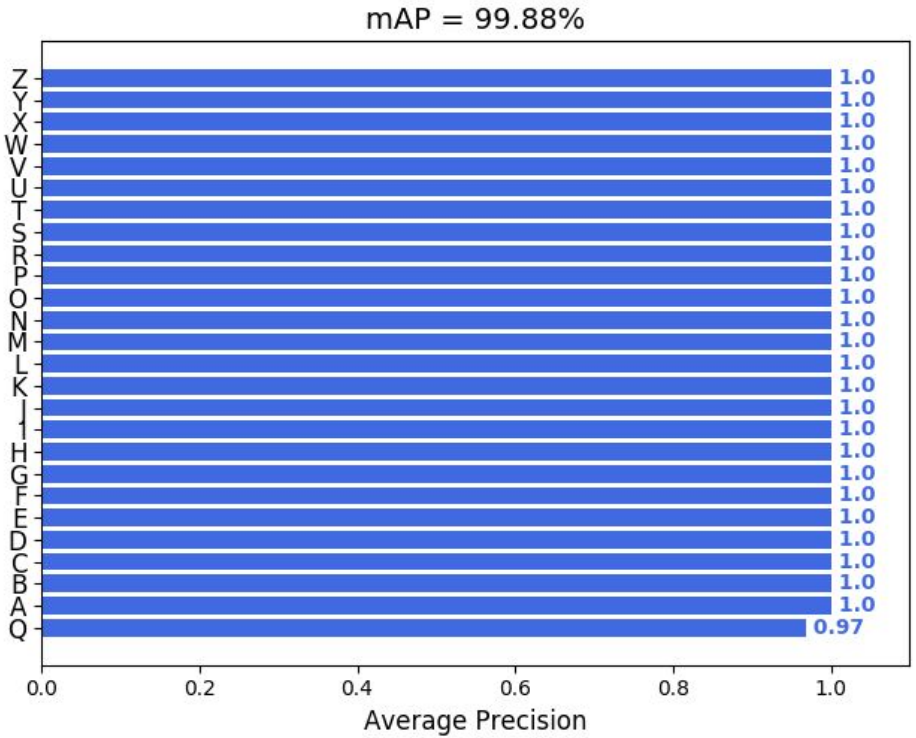
Network Configuration & Training

- Network configurations
  - assign 93 filters to all convolutional layers before every YOLO layer
  - Batch size of 64 with 16 subdivisions

3. Test & Results

Tests

1. Evaluate Mean Average Precision (MaP)
    a. Compare detections with annotations
    b. Used 1,496 of samples not considered for training
2. Accuracy evaluation with Kappa statistic
    a. Compare detections from video frames to annotations
    b. Create confusion matrix
    c. Calculate Kappa value



Results - Mean Average Precision (MaP)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	71	0	0	0	0	0	0	0
B	0	577	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	329	0	0	0	0	0	0	0	0	0	0	0	101	0	0	0	0	0	0	0	0	10	0	0
D	0	0	0	481	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	595	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	597	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	218	169	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	352	0	0	133	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	109	158	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0
J	0	0	0	0	0	0	0	0	133	441	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	1	0	0	0	0	11	123	225	4	0	0	0	0	0	2	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	604	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	41	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	138	0	0	0	0	0	0	0
N	6	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	44	36	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	566	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	3	0	0	0	0	6	0	0	0	0	48	62	0	0	5	0	0	0	0	0	11
Q	0	0	0	0	0	0	0	1	0	0	0	0	84	0	0	0	292	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	649	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	427	0	0	0	0	0	0	0
T	10	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	206	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	442	0	0	0	0	0
V	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	339	1	0	0	0
W	0	0	0	0	0	91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	219	0	0	0
X	0	0	0	0	0	0	0	0	10	4	0	0	0	0	0	0	0	3	0	0	0	0	0	152	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	542	0
Z	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85
	Kappa Value = 0.8382																									

Results - Kappa Statistic

https://www.youtube.com/watch?v=a8lxrEJuxxs&feature=youtu.be

Result - LARSIP Translation

4. Limitations

Dataset limitations

Limitations

Number of Samples

Amount of samples needs to be increased for better MaP

Complexity of Samples

Different backgrounds,distances, and hands must be considered

Expected Result

A more robust and adaptive hand detection and classification network

5. Future Work & Conclusion

Future Work

- Dataset expansion
- Gesture tracking
- Facial Recognition for contextual information
- Deployment on mobile devices

Conclusion

- Satisfactory results in terms of detection and classification
- Managed to detect and classify in speeds up to 30 fps
- Deployment of a mobile hand sign translation system is now feasible.

THANKS!

Any questions?

You can find me at:

- Juan.figueroa17@upr.edu
- LARSIP CID - F219