Multi-sensor System for Driver's Hand-Gesture Recognition

Pavlo Molchanov, Shalini Gupta,
Kihwan Kim, and Kari Pulli

Driver distraction

Current interfaces in cars distract drivers from the road

Driver distraction

(http://www.softkinetic.com)

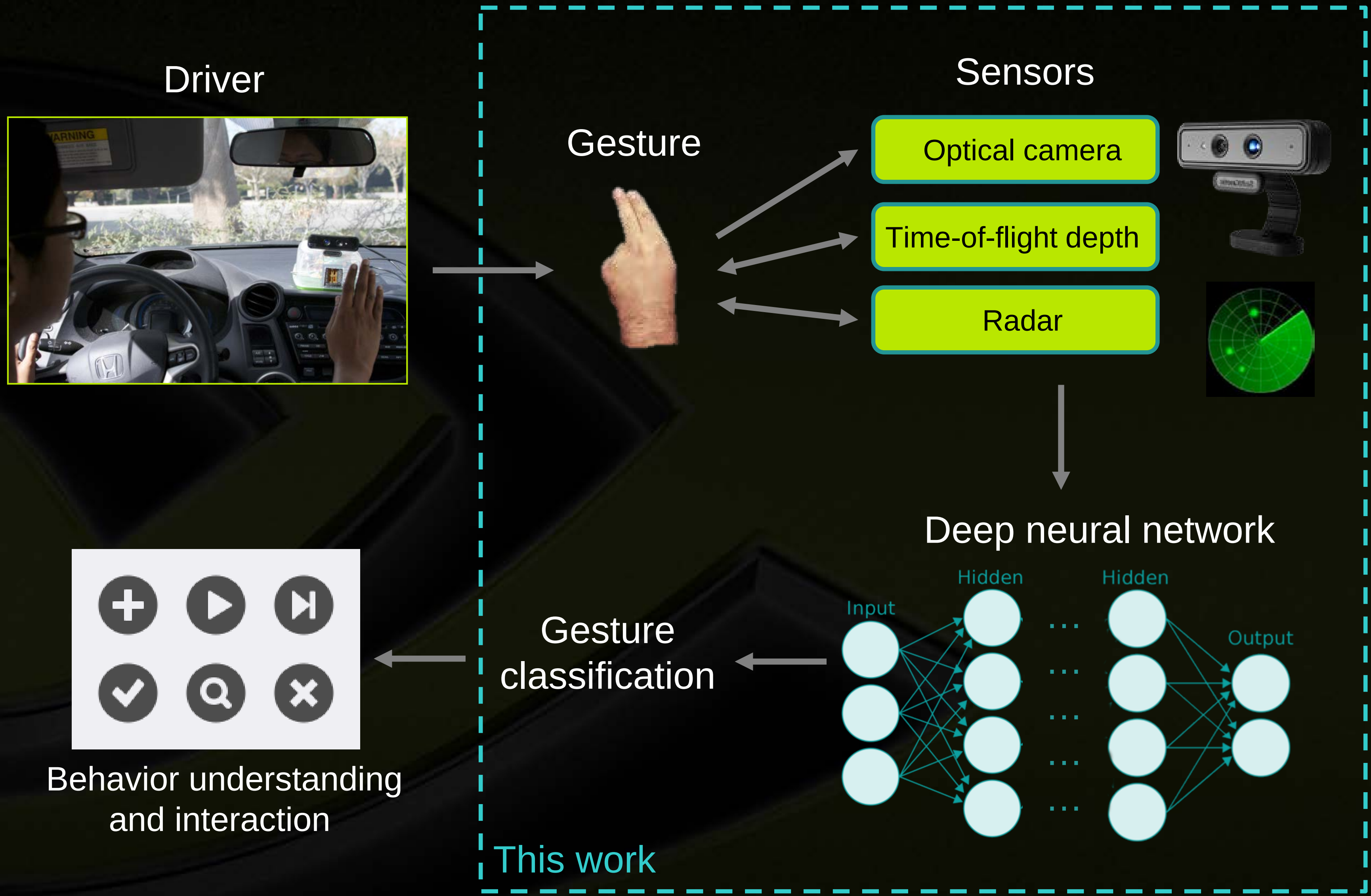
Touchless interfaces will help keep drivers' attention on the road

Existing work

- Neverova et al., *ECCV ChaLearn Workshop* 2014.
  - RGBD and upper body skeletal pose
  - Deep neural networks
  - Indoors only

- Ohn-Bar and Trivedi, *IEEE Trans. ITS* 2014.
  - RGBD
  - HOG+HOG^2 and SVM classifier
  - In car during day and evening

Our solution



Why multi-sensor?

Color cameras

Day

Night

Why multi-sensor?

Color cameras

Day

Night

Commodity depth cameras

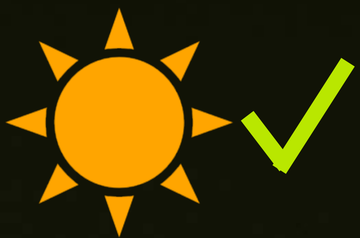
No sunlight

Sunlight

Radar sensor

All lighting conditions:

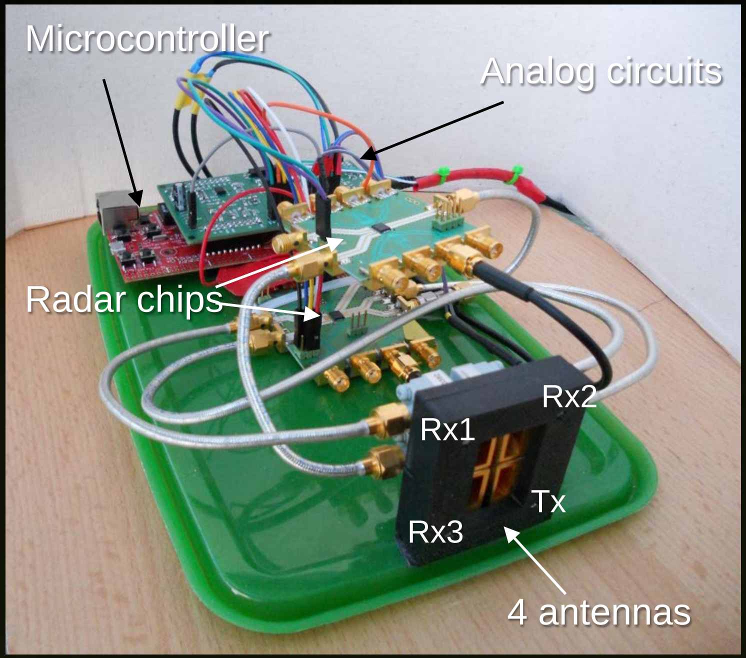
No interference from the Sun:



Direct measurements of local radial velocities (by Doppler shift):

Radar sensor

- Frequency Modulated Continuous wave (FMCW) radar architecture, 24 GHz
- Our design
- Molchanov et al., *IEEE Radar*, 2015.

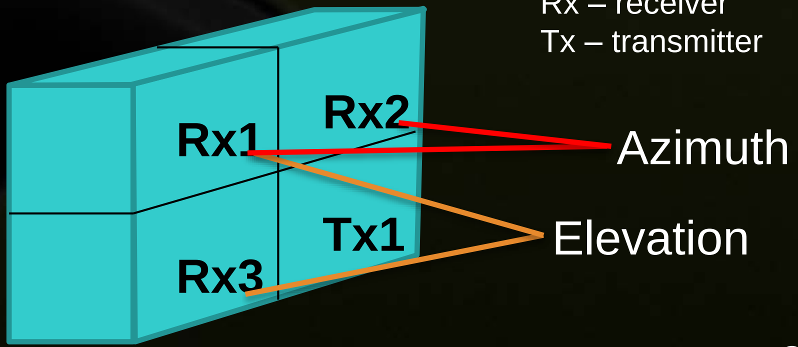


radar prototype

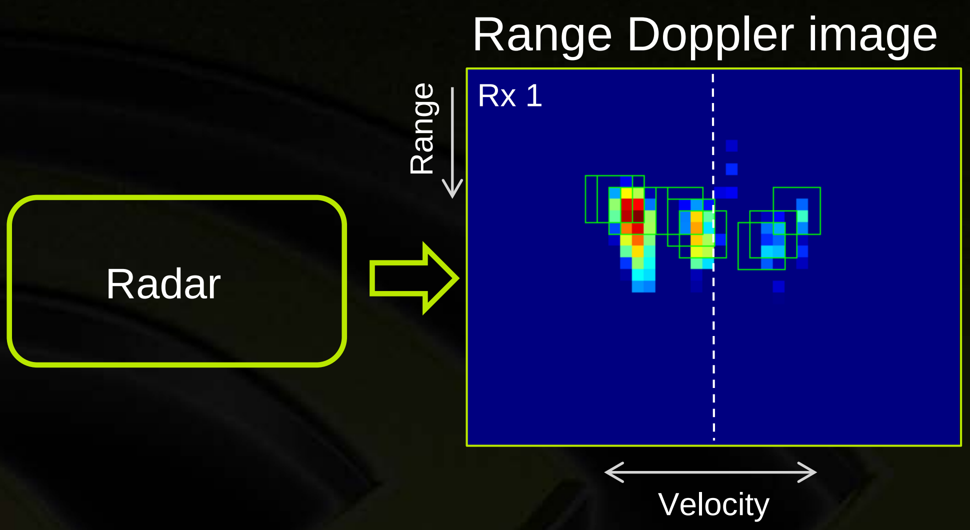
Radar can estimate:

- Range with resolution of 4 cm
- Radial velocity with resolution 0.04 m/s
- Angles of arriving (azimuth and elevation) are estimated for detected objects

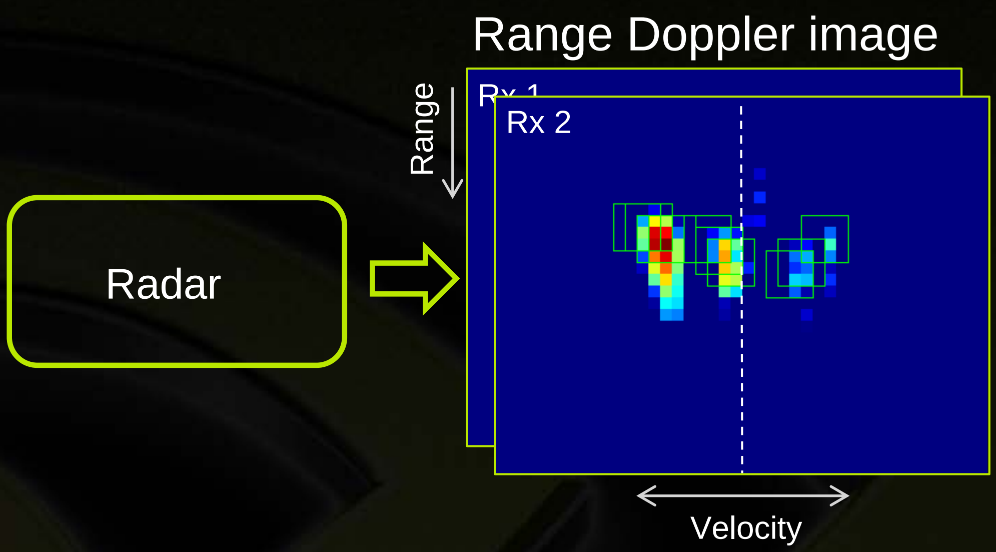
Rx – receiver
Tx – transmitter



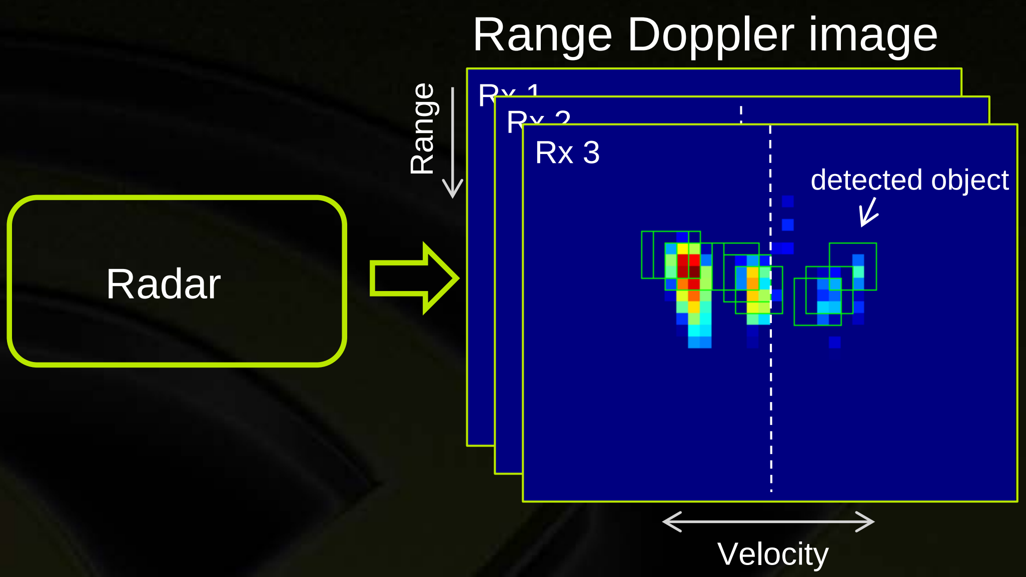
Radar pipeline



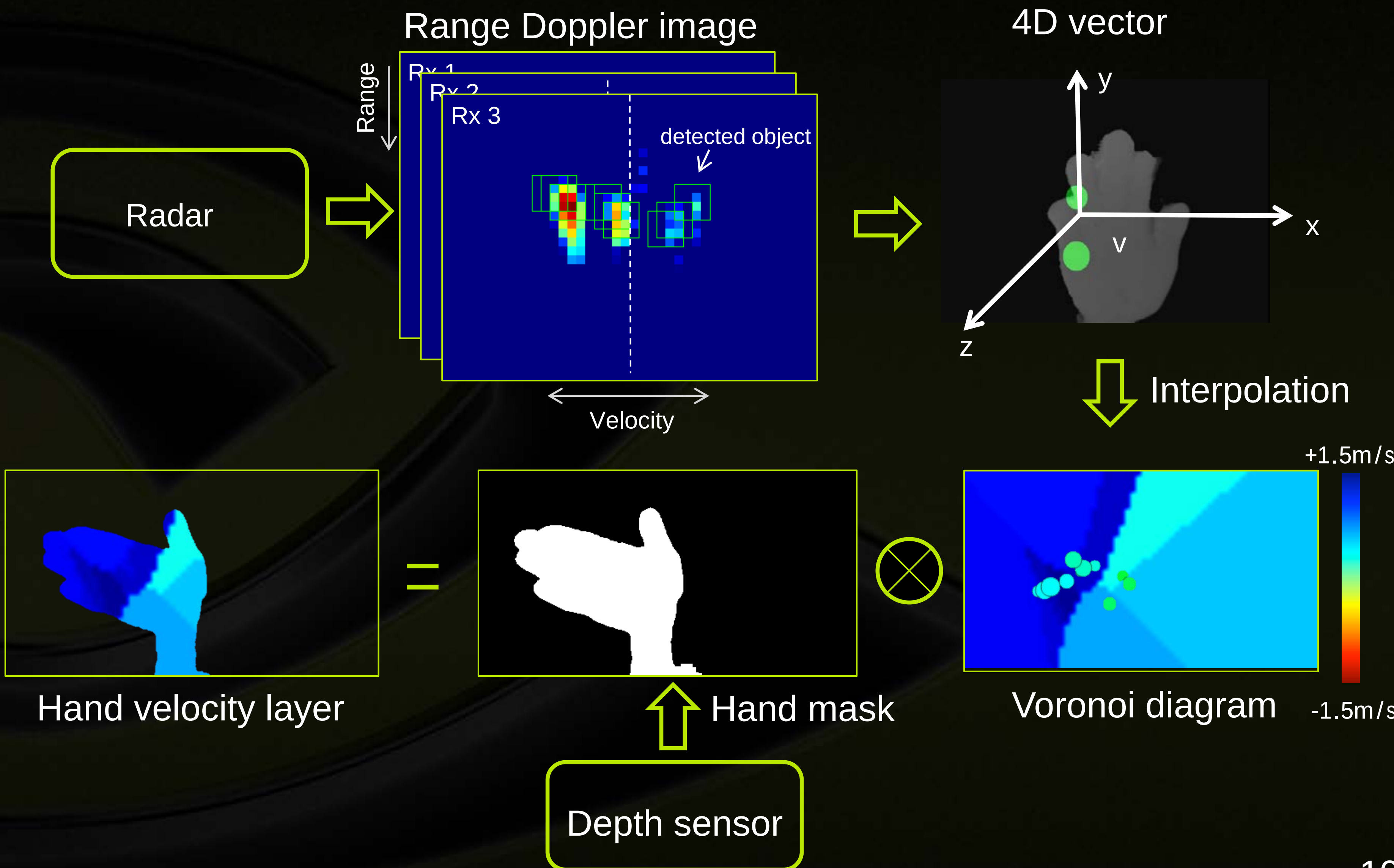
Radar pipeline



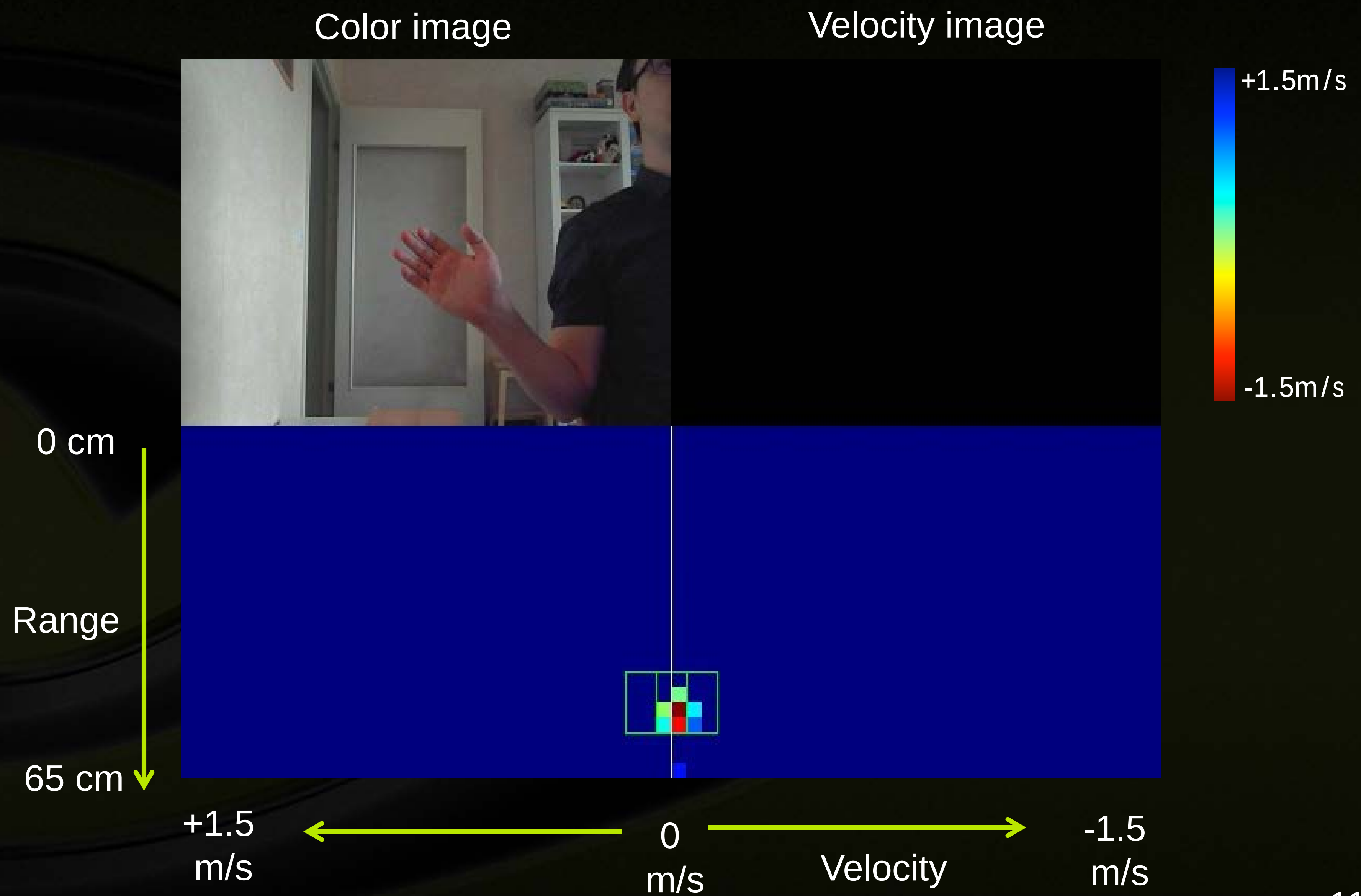
Radar pipeline



Radar pipeline

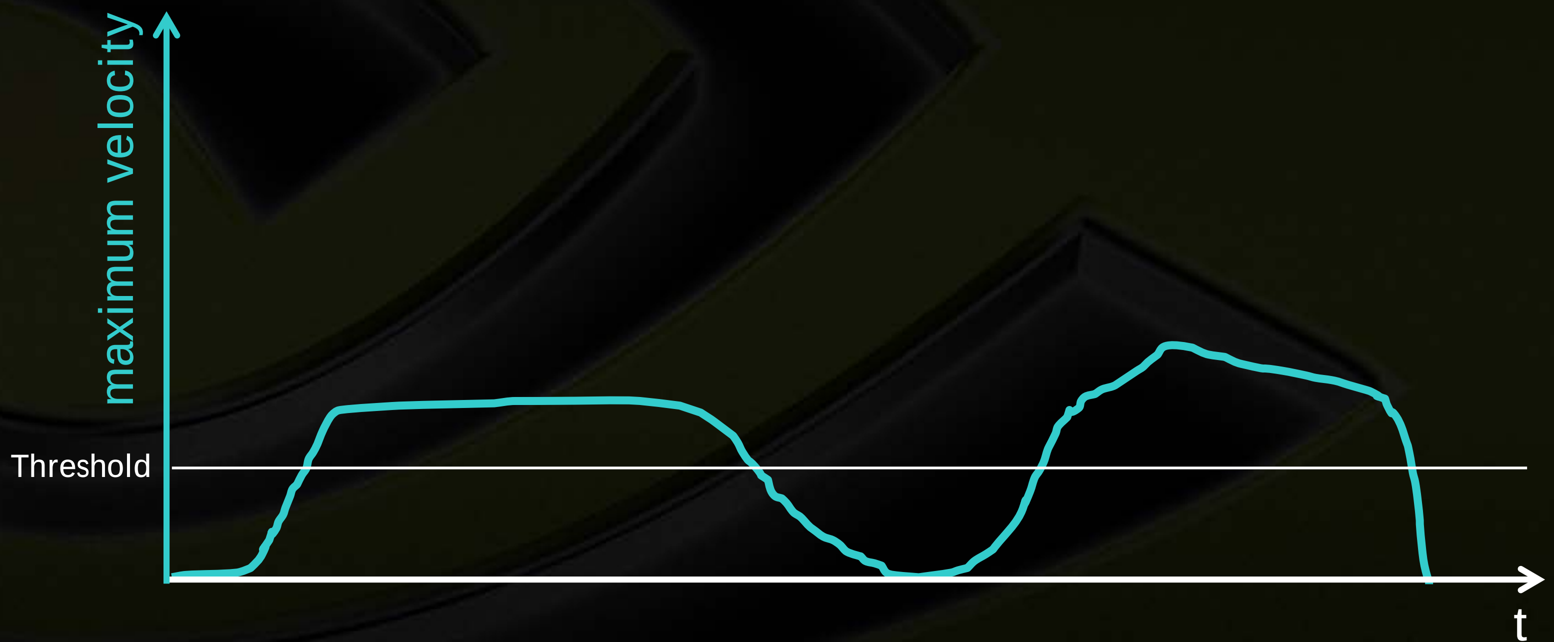


Radar sensor



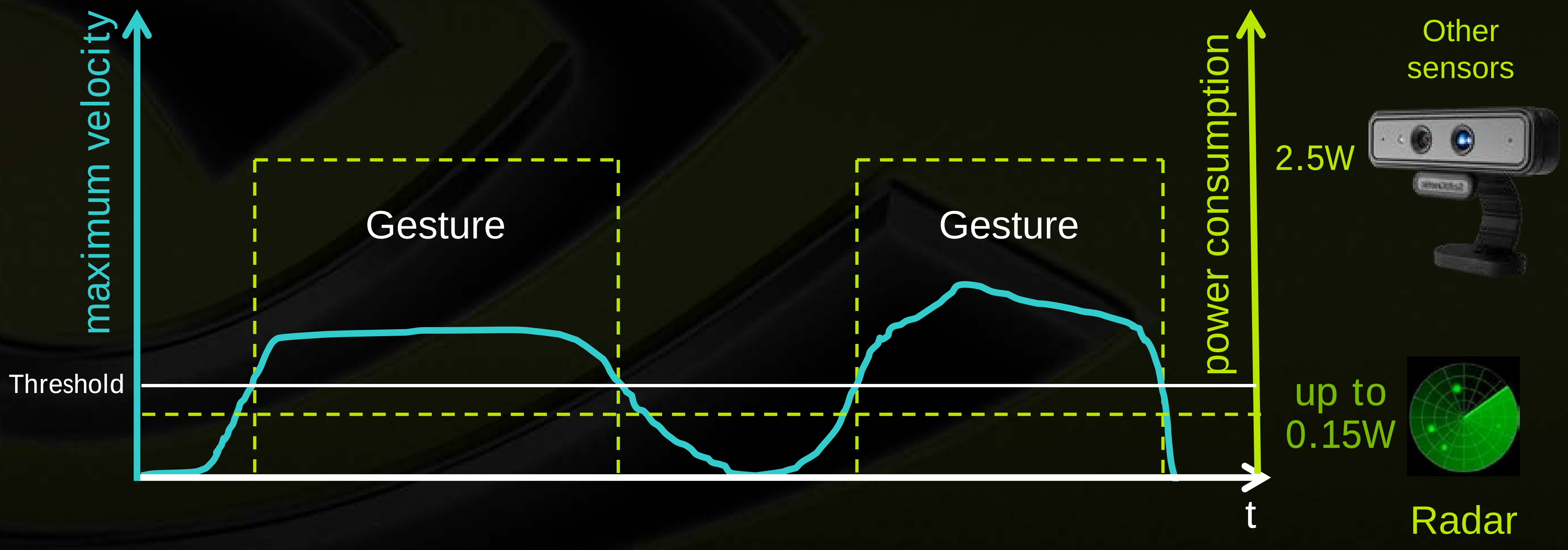
Segmentation

- Performed by radar
- Gesture detected when ***maximum velocity > threshold***
- Assumptions:
  - Hand stationary between gestures
  - Gesture duration 0.3 - 3s



Power efficiency

- Concept: High power sensors can be switched ON only during gesture detected by radar
- 16x power efficiency



Classification

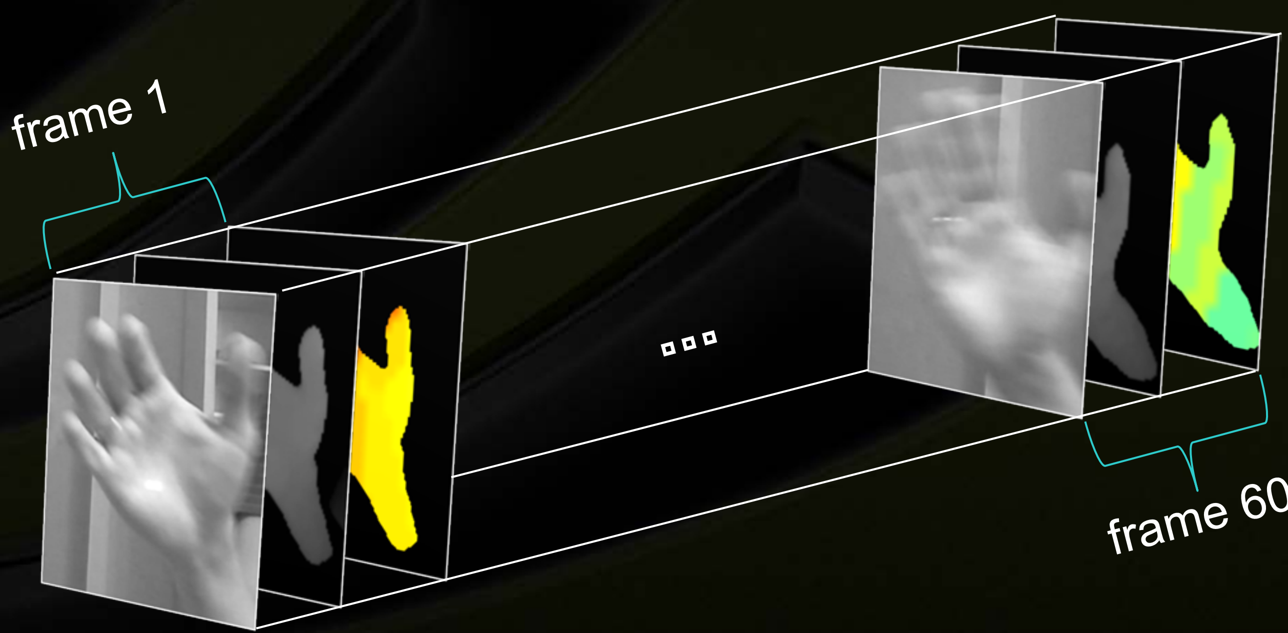
- Each frame consists of 3 channels: Intensity, Depth and Velocity
- Each channel is down sampled 32x32 pixels

Classification

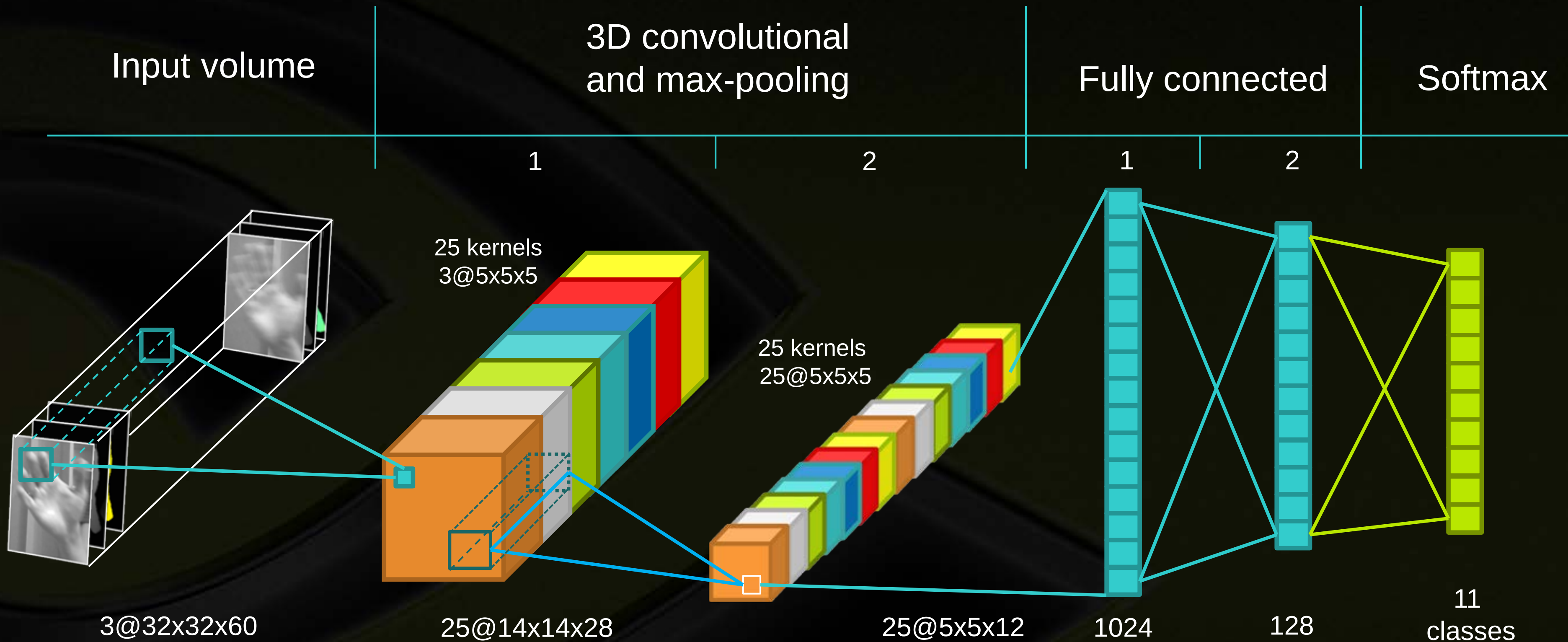
- Each frame consists of 3 channels: Intensity, Depth and Velocity
- Each channel is down sampled 32x32 pixels
- Segmented gesture is interpolated to 60 frames

Classification

- Each frame consists of 3 channels: Intensity, Depth and Velocity
- Each channel is down sampled 32x32 pixels
- Segmented gesture is interpolated to 60 frames



Gesture classifier



3D convolutional neural network

Data collection

- 10 experiments (sessions)
- 10 gesture types + random gesture
- 10-20 repetitions
- 3 persons

- Set-ups:
  - Indoor simulator
  - Outdoor, parked car

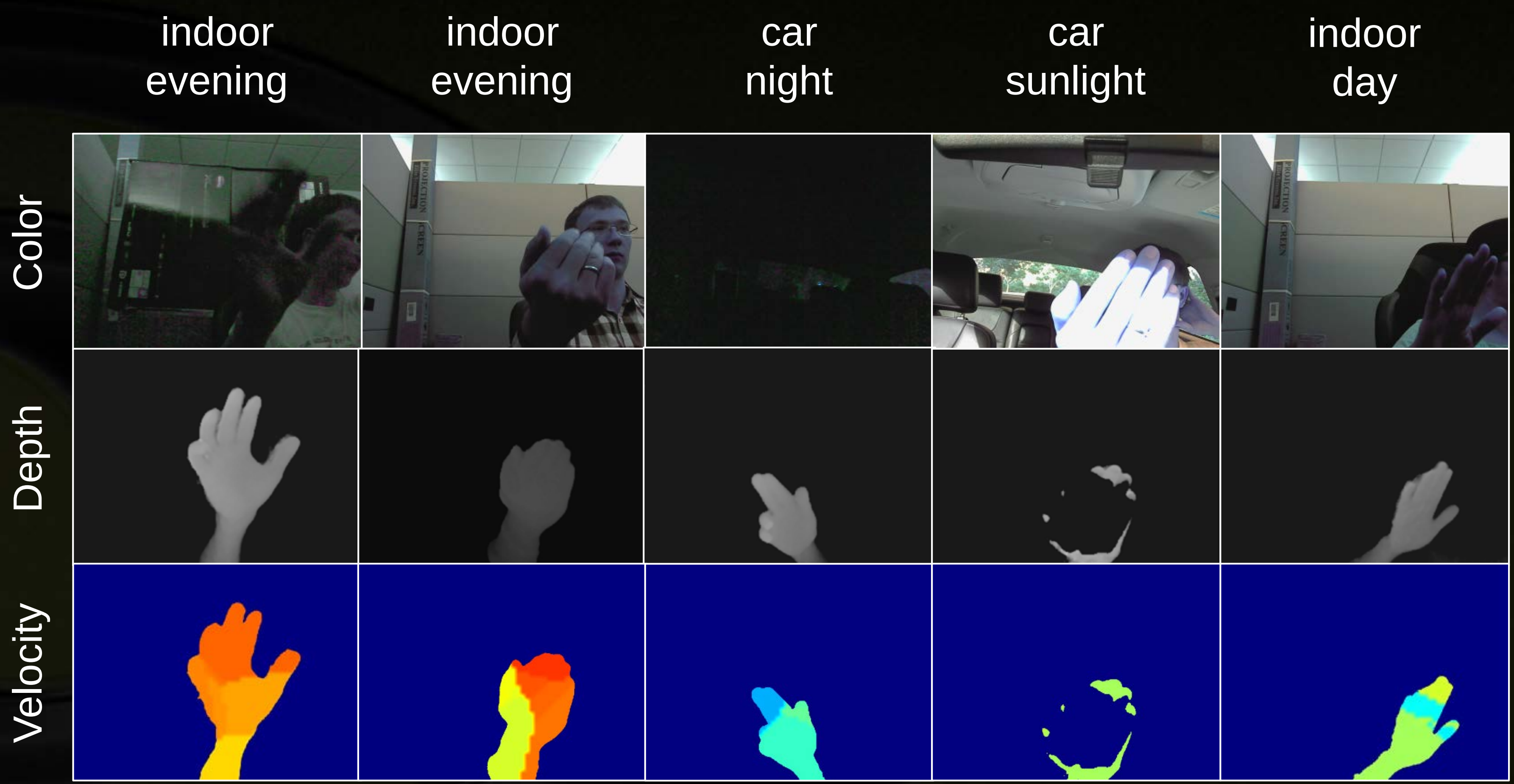
- Lighting conditions:
  - Day, shadow
  - Day, Sun
  - Evening
  - Night

- Total 1714 gestures

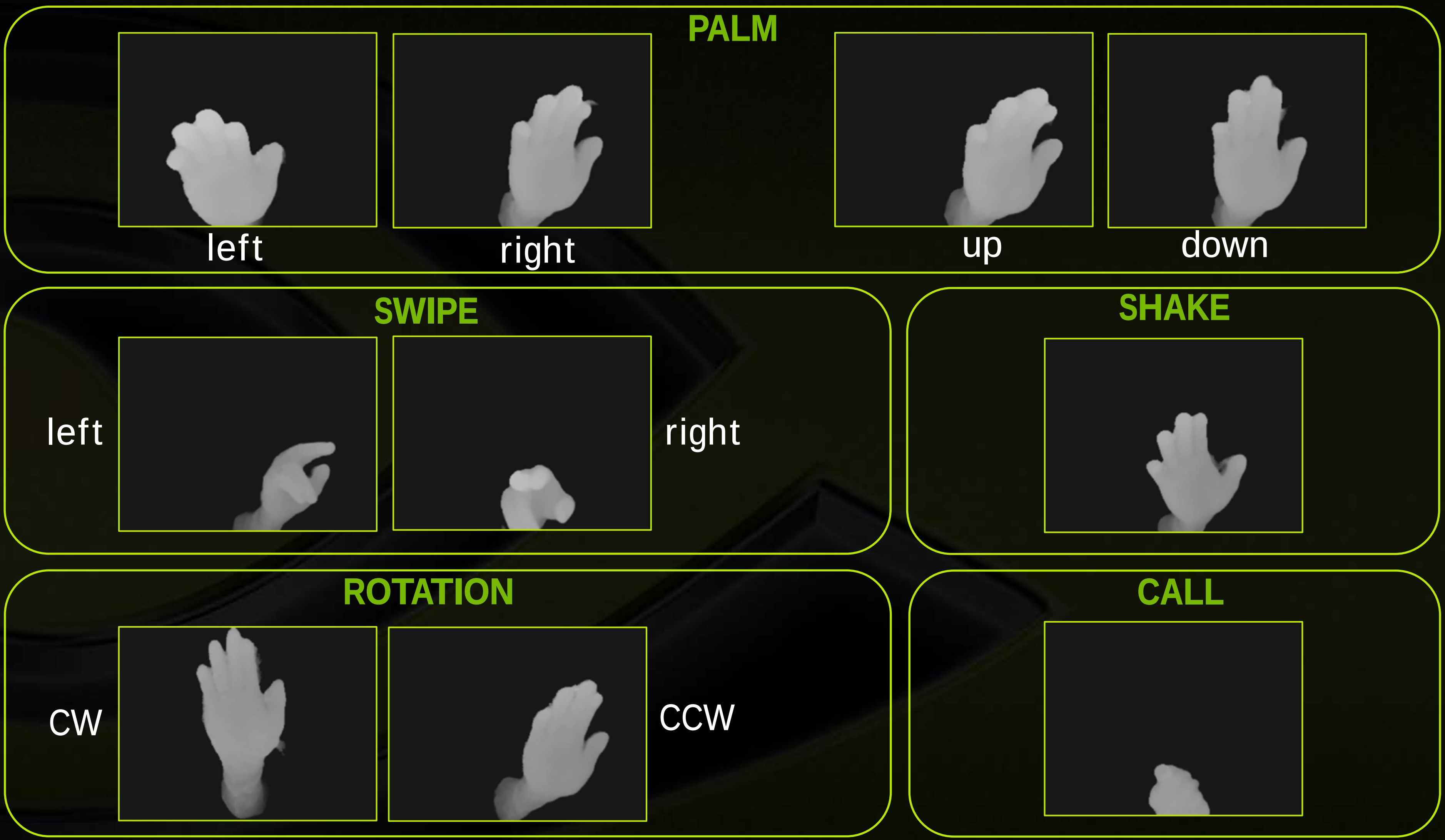
Indoor car simulator

Outdoor car

Data collection

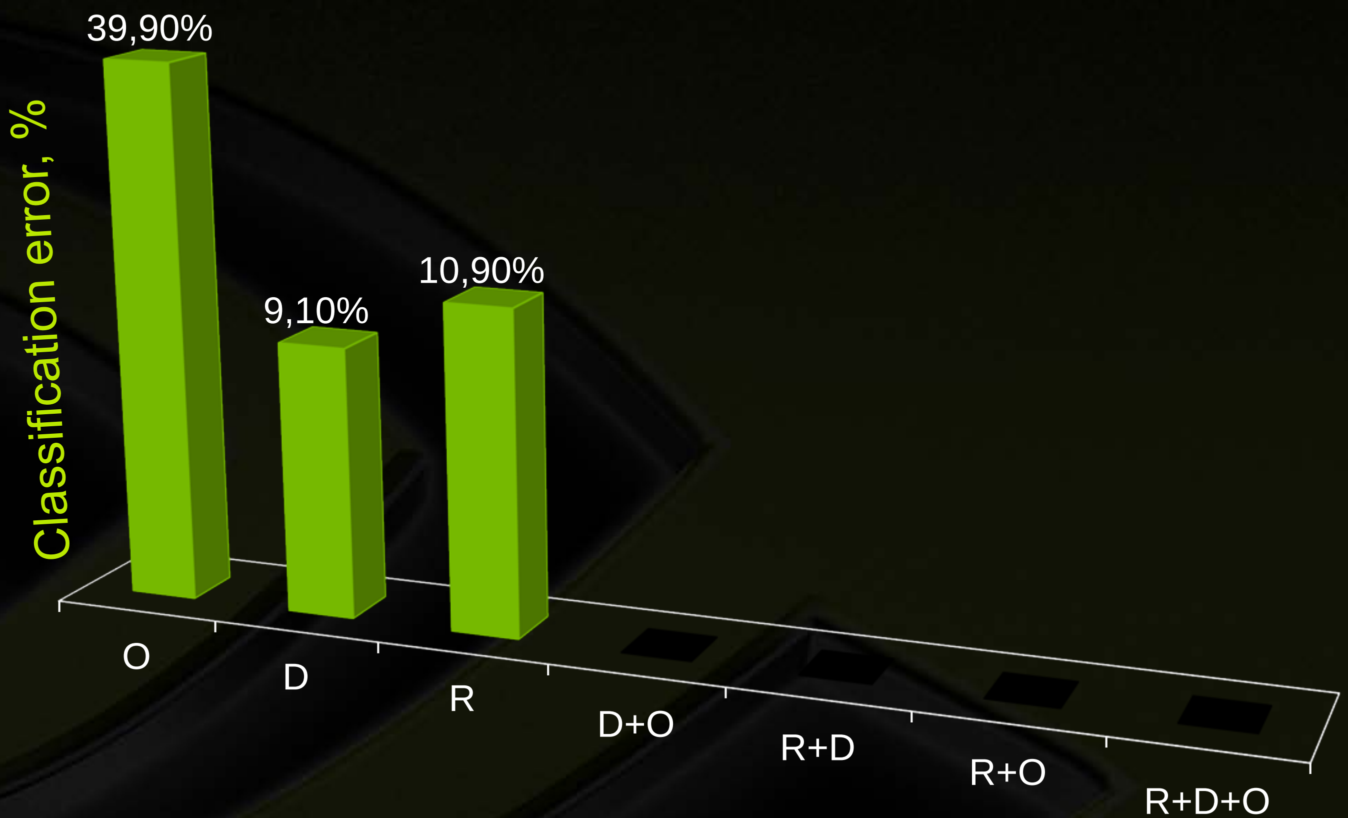


Gesture classes



Results

Leave one session out



D – depth O – optical R - radar

Results

Leave one session out

D – depth O – optical R - radar

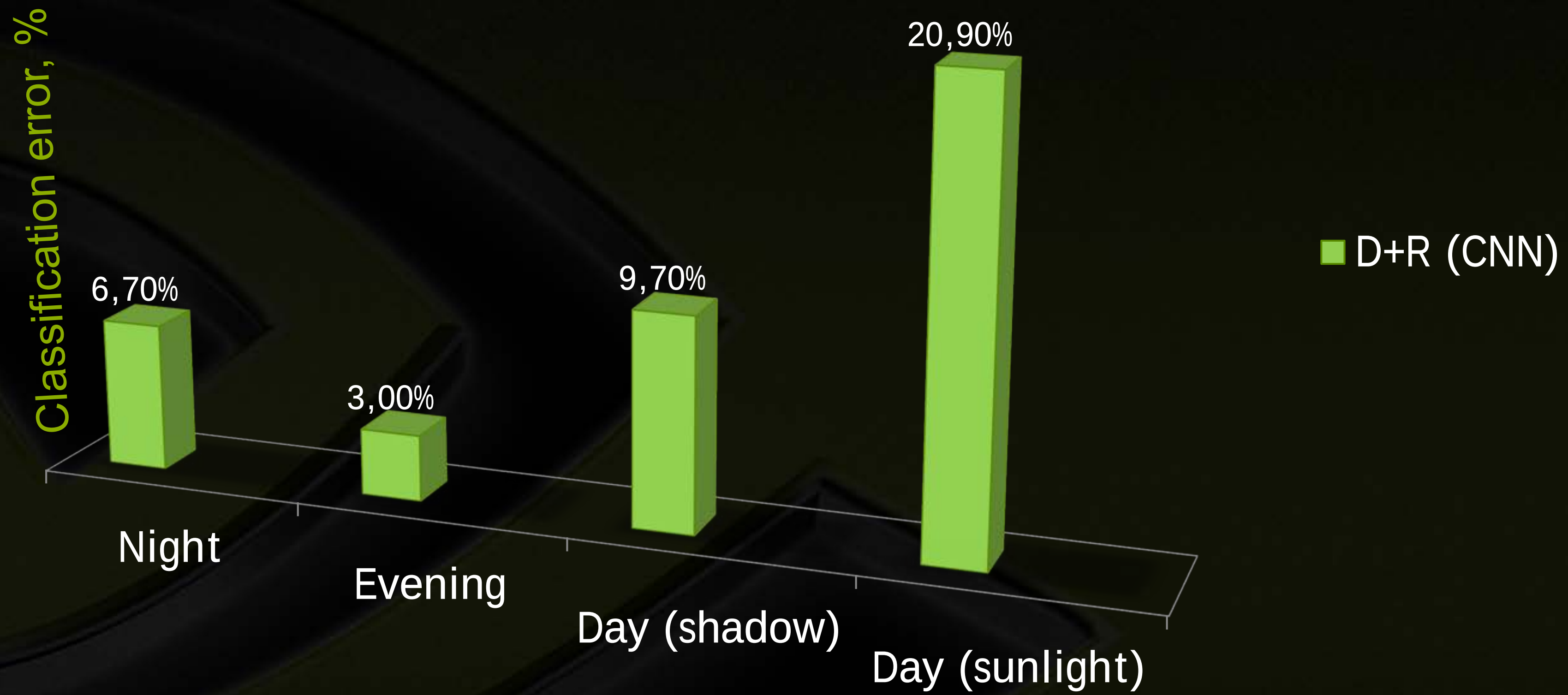
Results

Leave one session out

D – depth O – optical R - radar

Results

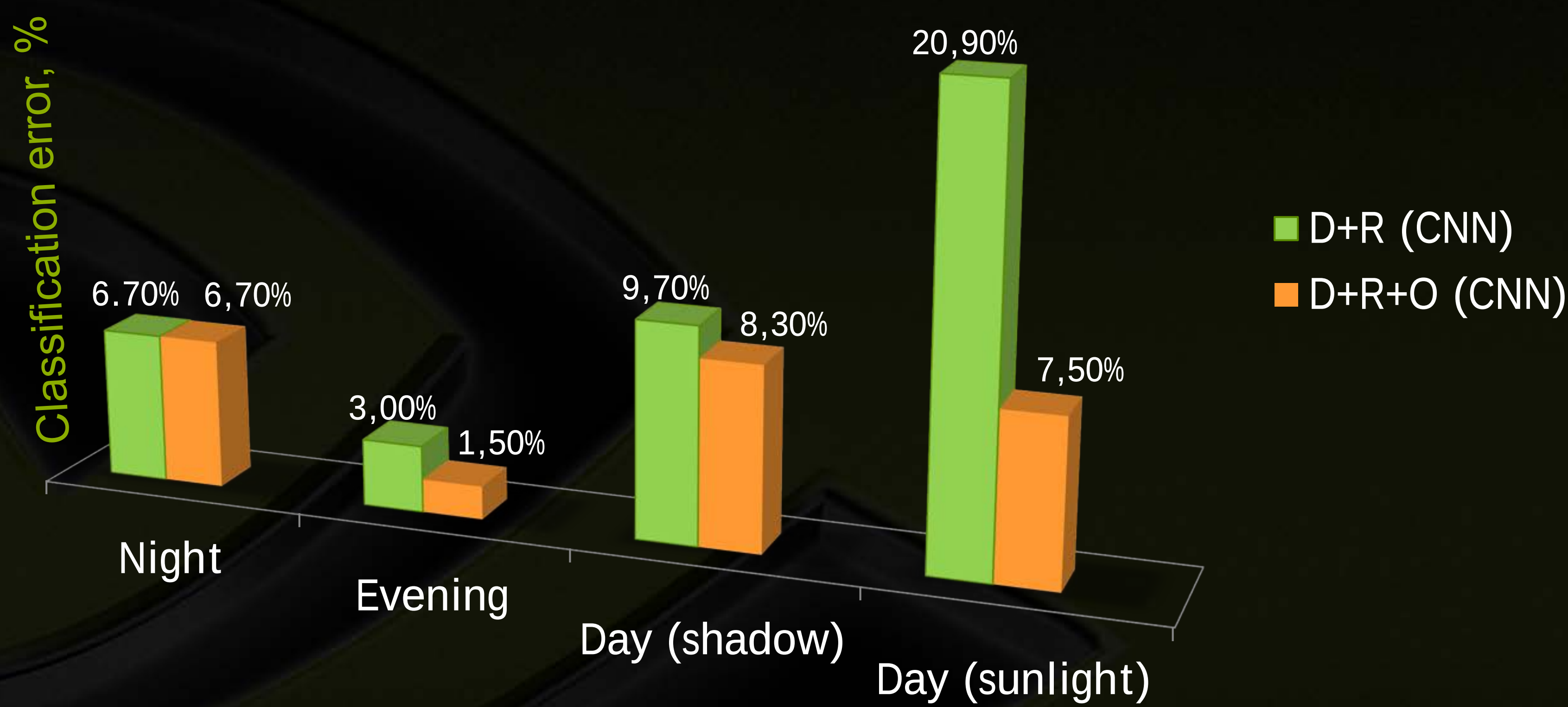
Leave one session out: different lighting conditions



D – depth O – optical R - radar

Results

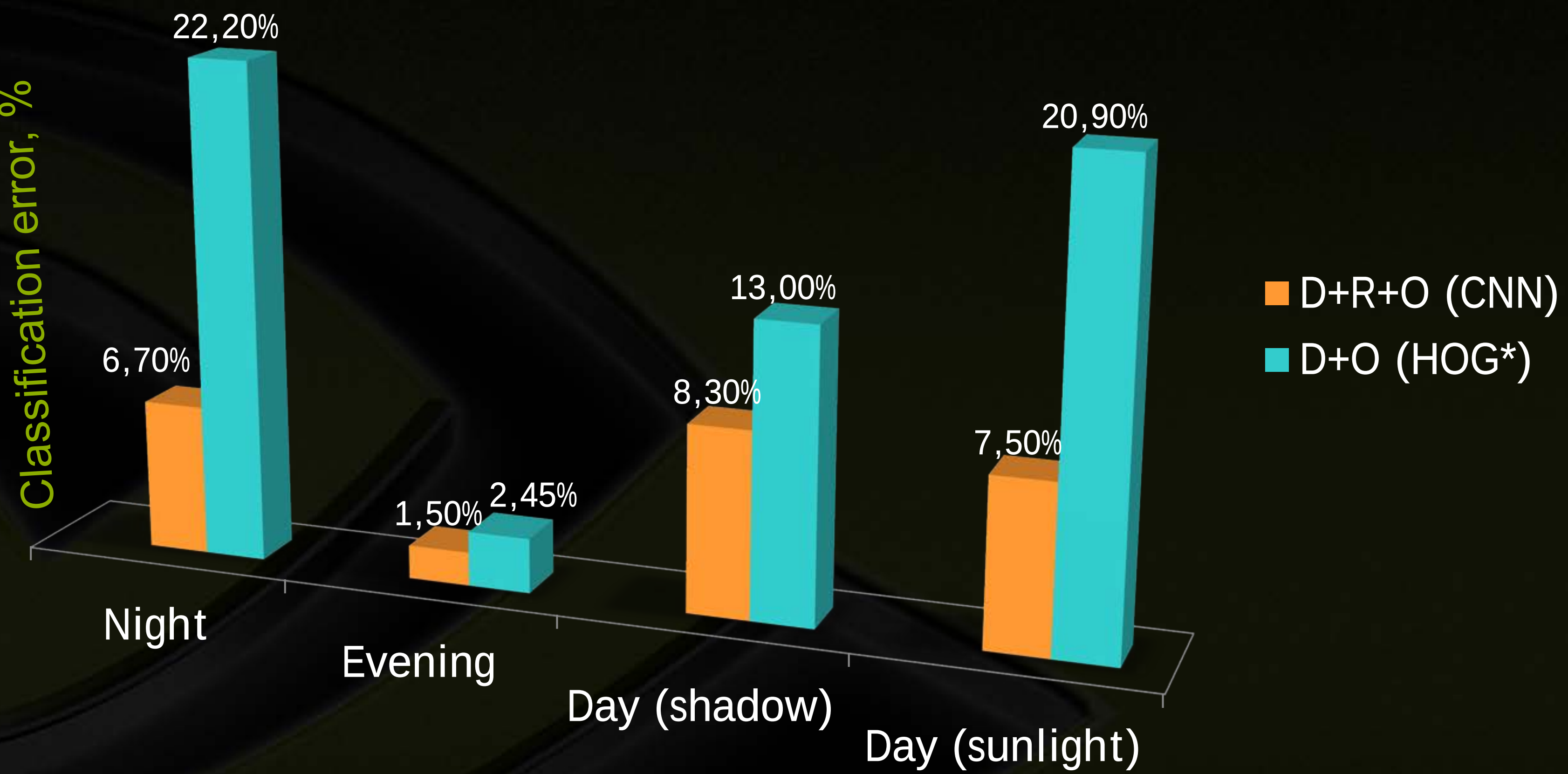
Leave one session out: different lighting conditions



D – depth O – optical R - radar

Results

Leave one session out: comparison with HOG features



D – depth O – optical R - radar

*Ohn-Bar and Trivedi, *IEEE Trans. on Intelligent Transportation Systems*, 2014.

Results

Classification error for leave one subject out classification

D+R+O (CNN)	D+O (HOG*)
24.90%	48.2%

D – depth O – optical R - radar

Solutions:

- Use more subjects to train
- Perform biometric registration of the system for a new user

*Ohn-Bar and Trivedi, *IEEE Trans. on Intelligent Transportation Systems*, 2014.

Demo

Gesture classification in parked car

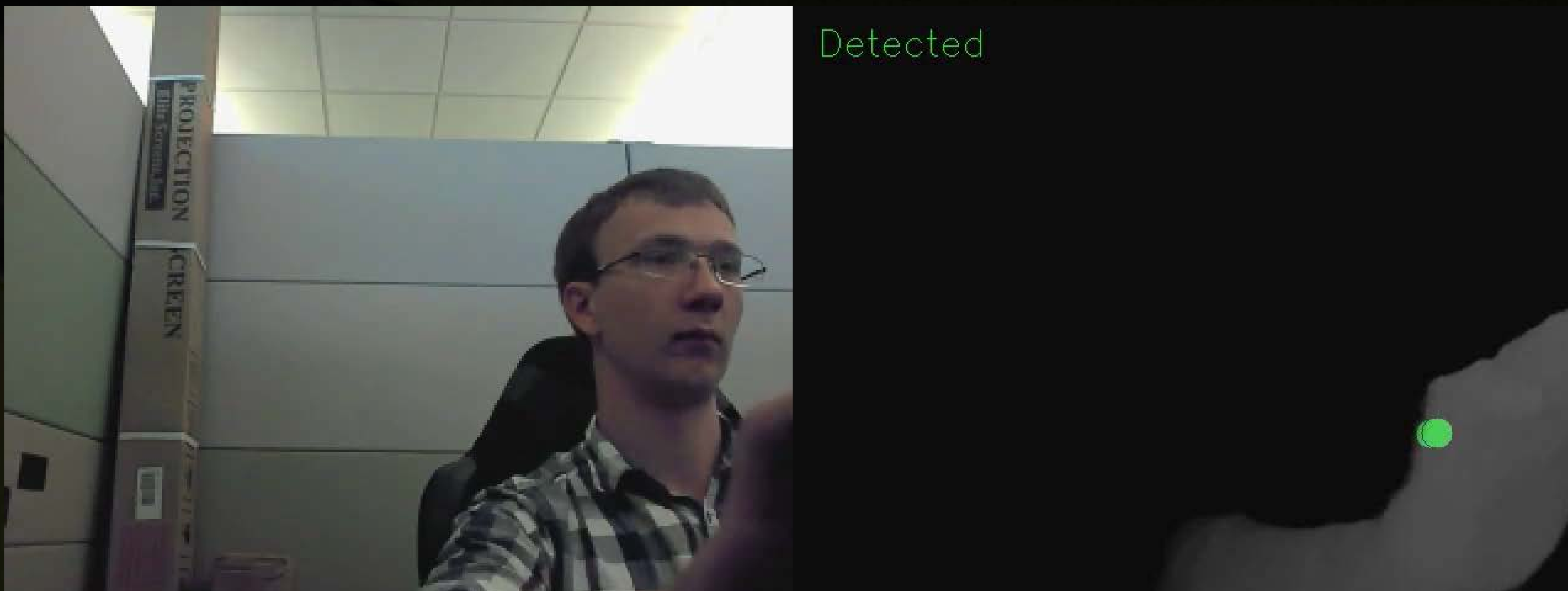
Conclusions

- A multi-sensor system improves accuracy and robustness to lighting
- CNN combines sensors effectively
- Proposed approach outperforms feature-based SOTA
- Using radar lowers power consumption, allows efficient gesture segmentation, and improves classification accuracy

THANK YOU
QUESTIONS?

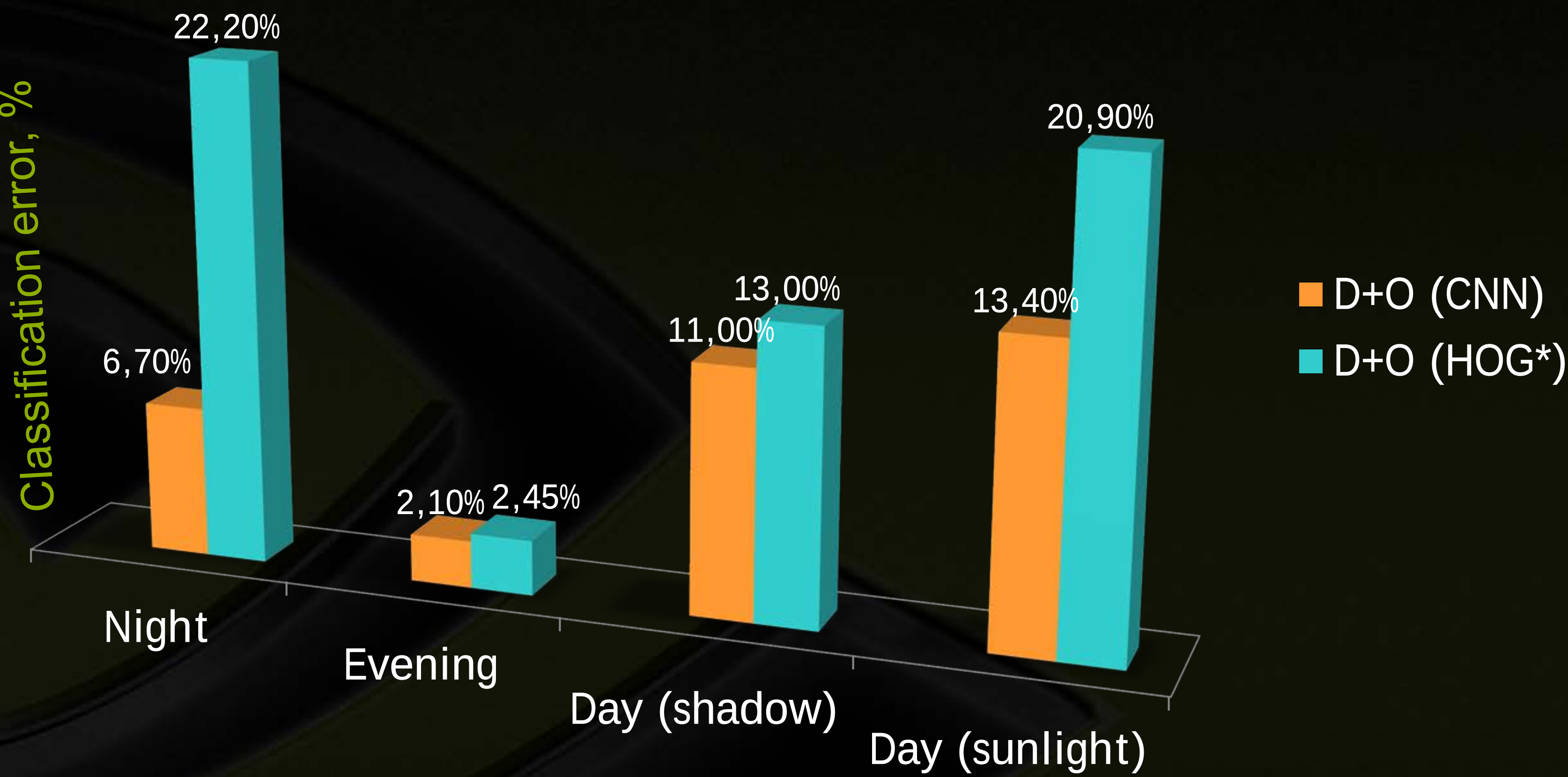
Demo

Gesture classification in simulator



Results

Leave one session out: comparison with HOG features

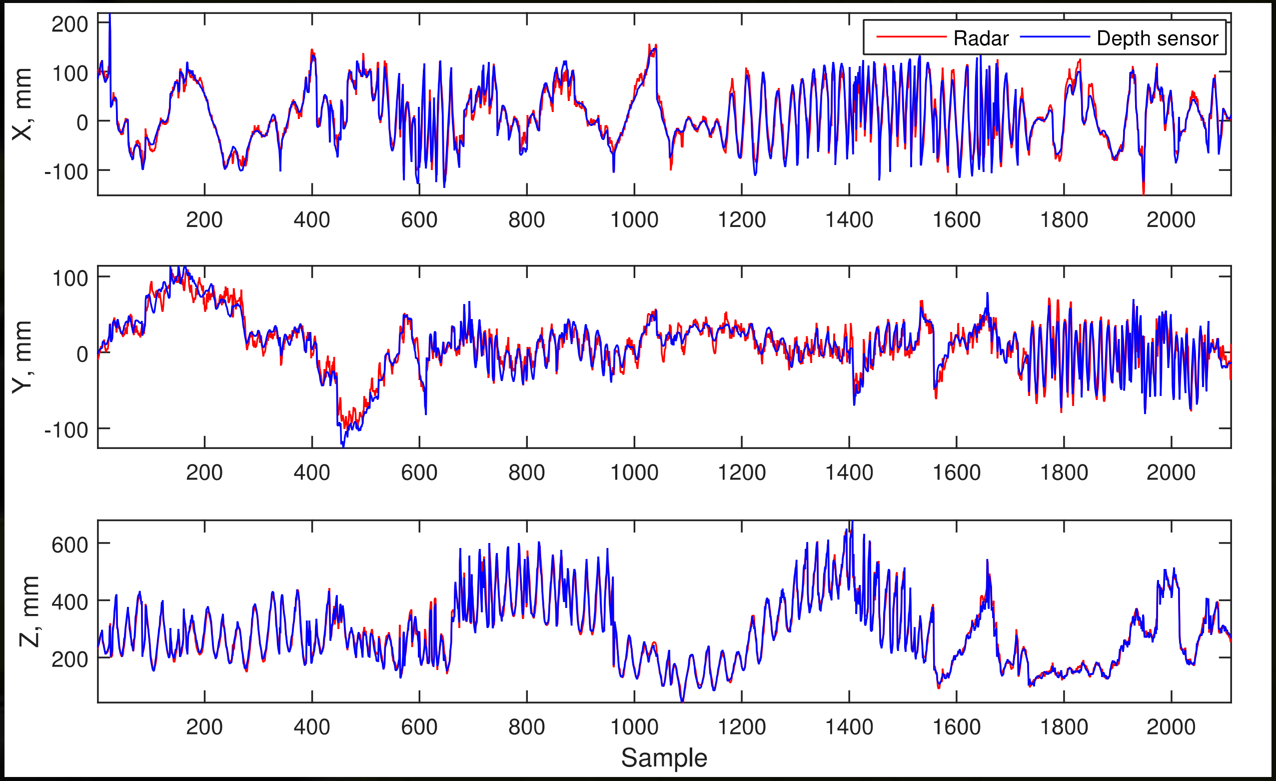


D – depth O – optical R - radar

*Ohn-Bar and Trivedi, *IEEE Trans. on Intelligent Transportation Systems*, 2014.

Calibration

Depth and Radar sensors are calibrated assuming a linear transformation model

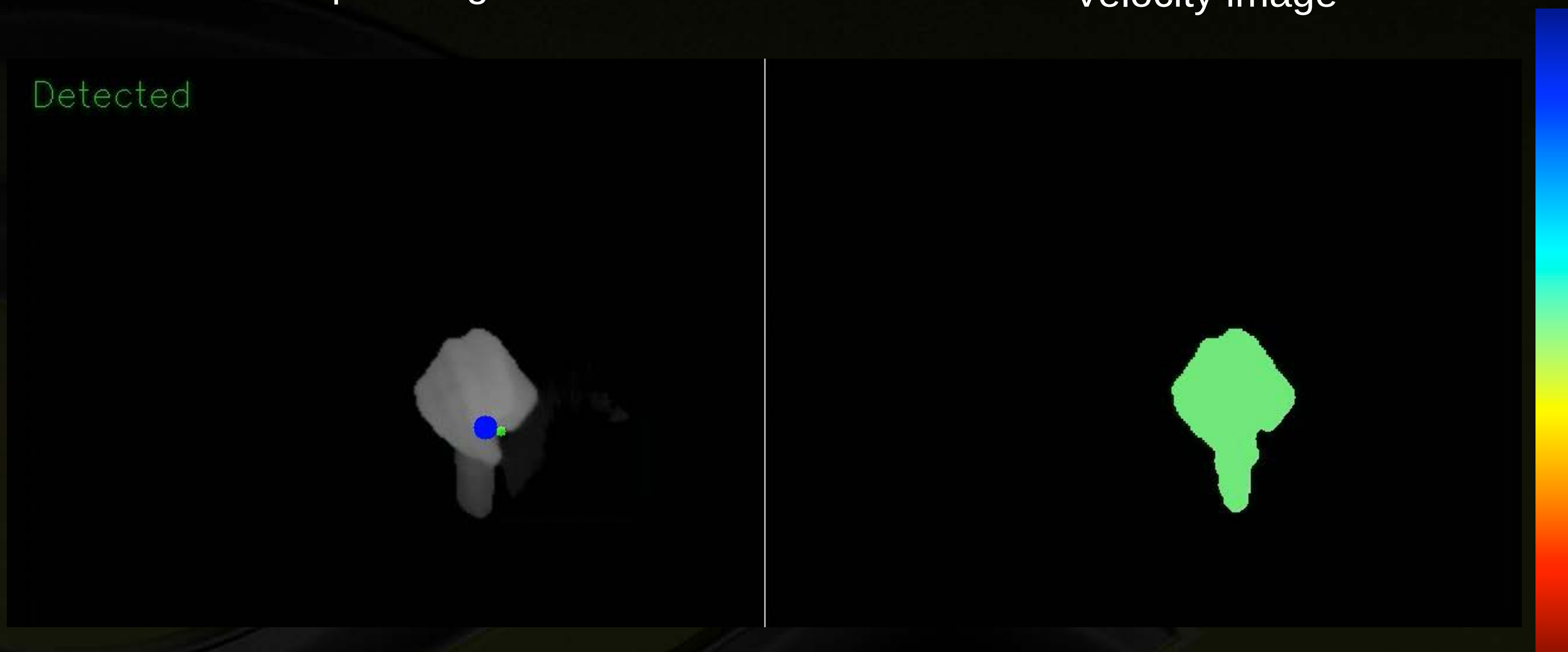


Radar pipeline

Depth image

Velocity image

+1.5m/s



-1.5m/s

Green circles – points detected by radar
Blue point – hand position estimated from the radar