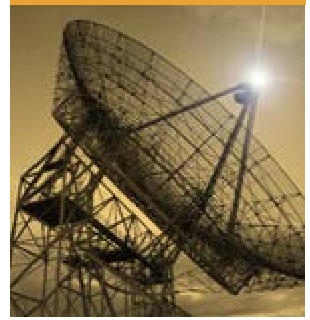
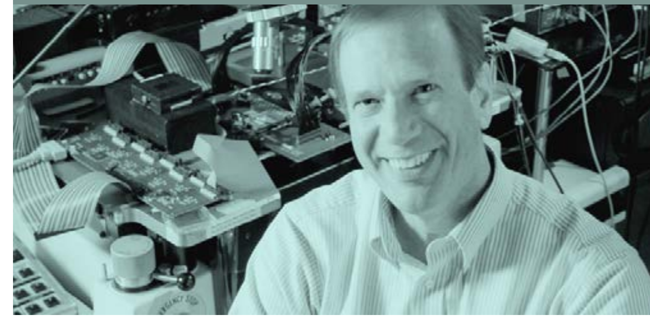
SRI International





Fifth International Symposium on **Naturalistic Driving Research**

DCode: A Comprehensive Automatic Coding System for Driver Behavior Analysis

FHWA Exploratory Advanced Research - Topic 2A

Amir Tamrakar, PI

Gregory Ho, Jihua Huang, David Salter, Avi Ziskind, Chenyang Zhang, Yin Xia, Yilin Song , Wei Li

SRI International, Princeton, NJ

U.S. Department of Transportation
Federal Highway Administration

August 30, 2016
Site: VTTI, Blacksburg, VA

The Need

SHRP2
STRATEGIC HIGHWAY RESEARCH PROGRAM

- Naturalistic Driving Study (NDS) under the SHRP2 program
  - Collected normal driving behavior data
    - 3,400+ drivers
    - 5,400,000+ Trip
    - ~1 Million hours of video data + other metadata

- There is way too much data for manual coding!

- FHWA EAR 2A program (2014-2016) was created to help explore the feasibility of automated coding of SHRP2 NDS
  - Explore existing technologies
  - Develop new technologies

DCode: Technology Concept

- **Goal:** **A**ssist in the automatic coding of features relevant to safety researchers interested in using the SHRP2 NDS data using Computer Vision techniques.

- A comprehensive driving behavior study will need to take into account not only the actions and behaviors of the driver but also the "***context***" in which those actions are performed
  - *Context* = everything external to the driver's person

Contextual Features:

Factors Outside the Vehicle:
Traffic condition,
Weather conditions,
Road conditions,
Actions of and signals from Pedestrians, Bicycles, Vehicles,
Traffic lights, Road signs.

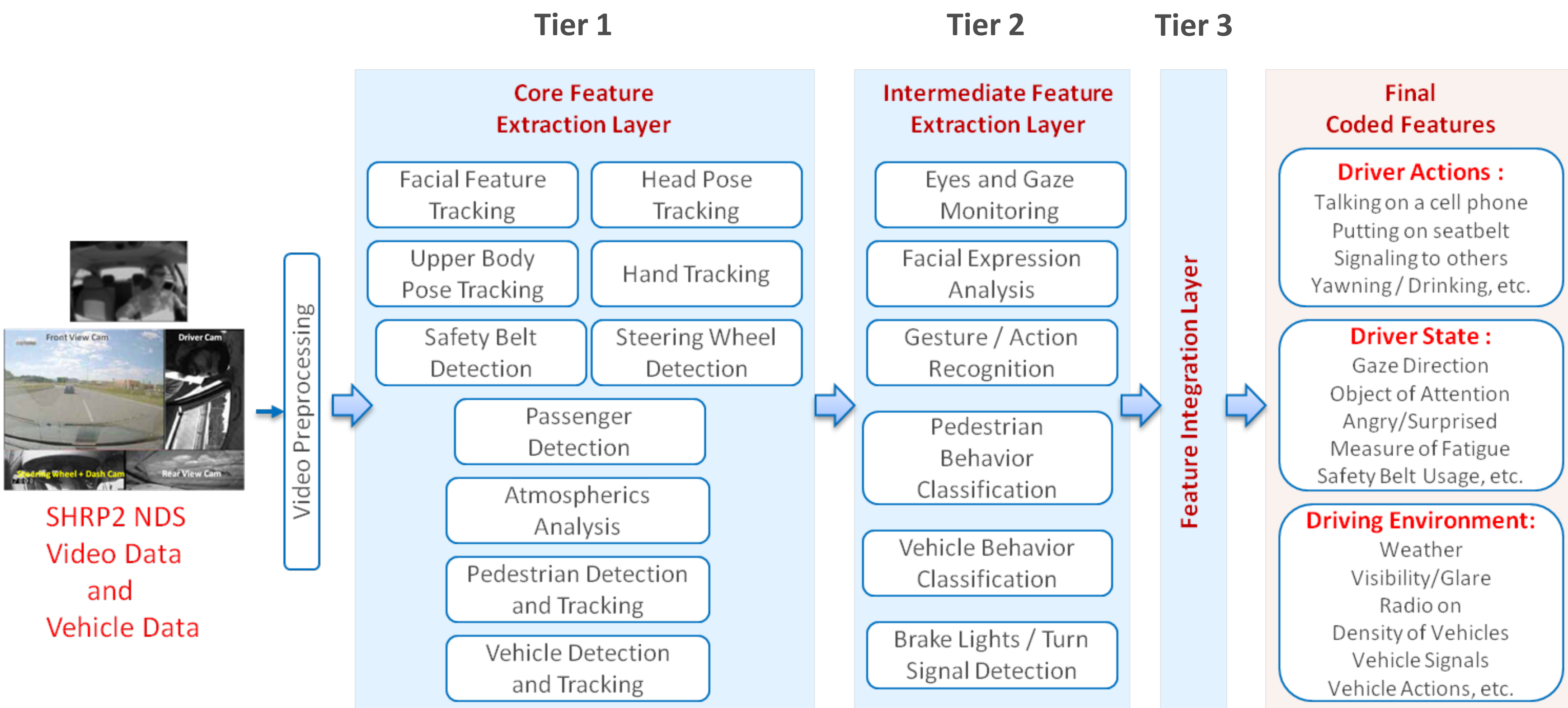
Factors Inside the Vehicle:
Passengers,
Passenger-caused distractions,
Sat. Nav., Radio,
Cell phones, Travel Mugs
Gadget-caused distractions

Driver Features:

Driver State:
Head Pose,
Gaze, Eye Blinks,
Mouth Movement,
Facial Expressions,
Hands,

Driver Actions:
Gestures,
Actions

DCode: Technical Plan Overview



- Lane trackers,
- Accelerometers,
- GPS,
- Cell phone records,
- Vehicle operation data

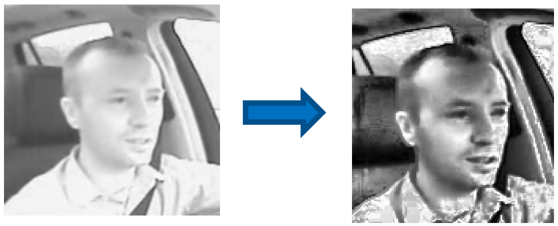
- Companion Roadway Information Data.

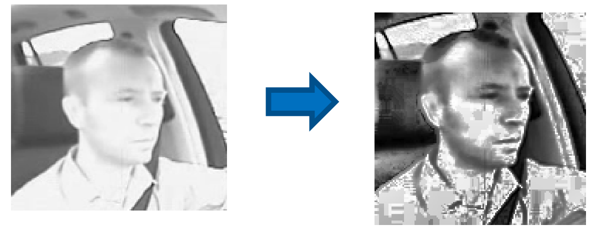
SHRP2 Dataset Automated Coding Challenges

- Unique challenges for computer vision algorithms
  - Very low resolution (240x356 wide FOV, 70x70 pixels on the face)
  - Heavy compression artifacts (gets worse with fast illumination changes)
  - Uncontrolled Illumination
    - High degree of influence from external factors
    - Extremely poor contrast (often completely saturated)
    - Fast lighting changes
  - Poor illumination for night time sessions
  - Camera viewpoints
    - Camera placed at an angle

SHRP2 Raw Video Data

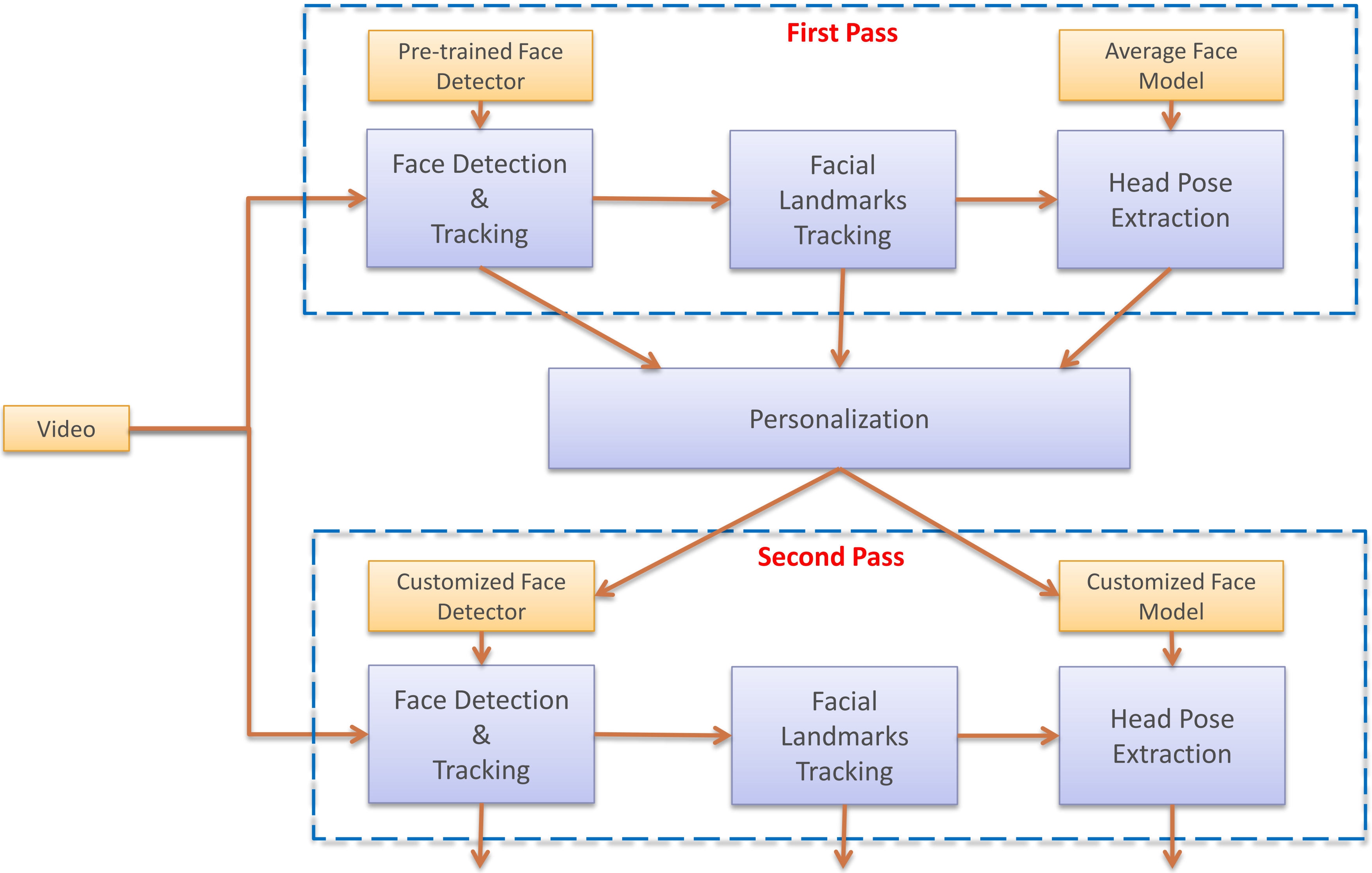
Core Feature: Driver's Face Detection and Tracking





Some video preprocessing

Our Approach to Customized Face Tracking



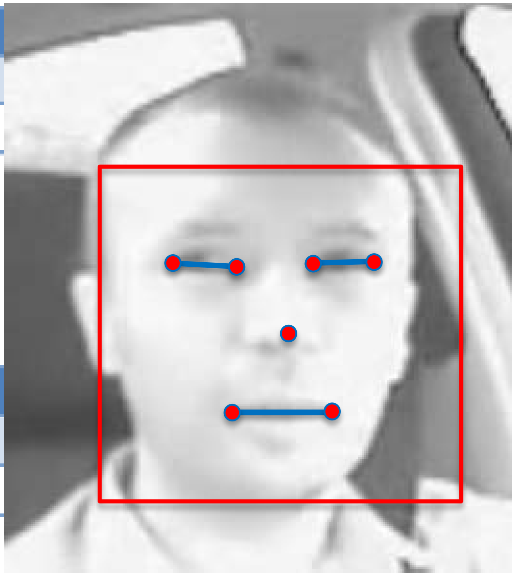
SHRP2 Data Annotations For Evaluation

- Our dataset is the SHRP2 24-car study (HPV dataset)
- We are using the metadata contained in VTTI's HPV mask public dataset (along with ORNL's Matlab scripts and data formats)
  - More than sufficient for evaluating driver tracking algorithms.
  - Annotated segments are harder than average because of prompted tasks.

44 videos -- 22 static trial vides and 22 dynamic trial videos

SHRP2 HPV
Hi-res Face Video
Dataset

General Statistics	
Total Duration of all the video	17.98 hrs
Total # of frames processed	970,847



SHRP2 HPV
Hi-res Face Video
Annotated Clips
Subset

General Statistics	
Total Duration of all the video	1.38 hrs
Total # of frames processed	74978

7% of all frames

Precision-Recall Curves for Face Detection

$$Bounding\ Box\ overlap\ ratio := \min\left(\frac{Area\ of\ Overlap}{Area\ of\ GT\ Bbox}, \frac{Area\ of\ Overlap}{Area\ of\ Detection\ Bbox}\right)$$

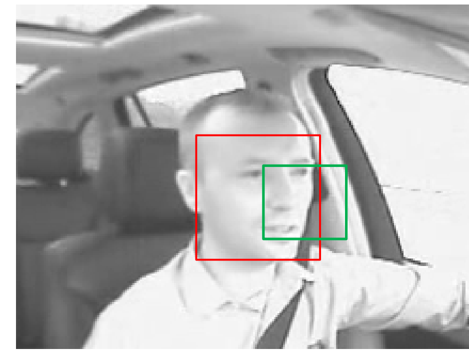
$$Precision\ Rate := \frac{\#of\ Faces\ that\ were\ correctly\ "detected"}{Total\ \#\ of\ Faces\ that\ were\ "detected"}$$

$$= \frac{\#\ of\ detections\ where\ bbox\ overlap > match\ threshold\ and\ score > threshold}{\#\ of\ detections\ where\ score > score\ threshold}$$

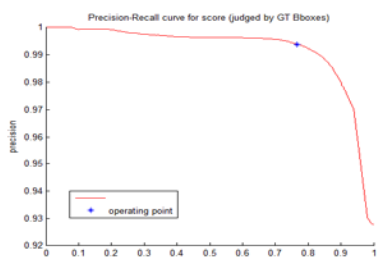
$$Recall\ Rate := \frac{\#of\ Faces\ that\ were\ "detected"}{Total\ \#\ of\ Faces\ in\ the\ data}$$

$$= \frac{\#\ of\ detections\ where\ score > score\ threshold}{Total\ \#\ of\ Faces\ in\ the\ data}$$

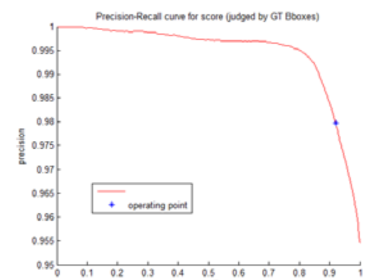
Overlap ratio = 0.92



Overlap ratio = 0.2



First Pass:
Recall = 79.58% and Precision is 99.26%.



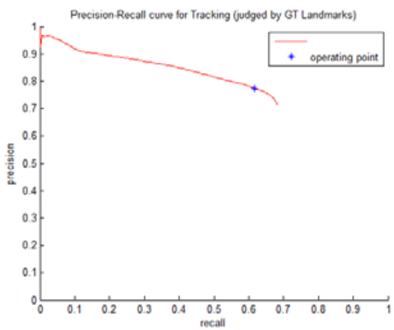
Second Pass:
Recall = 96.06% and Precision is 96.54%.

Bbox match threshold = 0.50

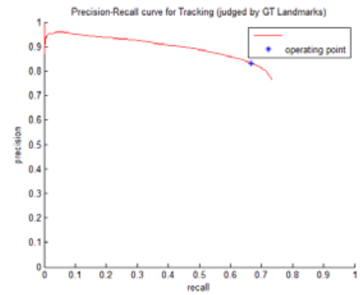
Precision-Recall Curves for Facial Landmark Tracking

$$Tracking\ Precision\ Rate: = \frac{\#of\ Faces\ that\ were\ accurately\ "tracked"}{Total\ \#\ of\ Faces\ that\ were\ "detected"} = \frac{\#\ of\ detections\ where\ tracking\ error < error\ threshold\ and\ score > threshold}{\#\ of\ detections\ where\ score > score\ threshold}$$

$$Recall\ Rate := \frac{\#of\ Faces\ that\ were\ "tracked"}{Total\ \#\ of\ Faces\ in\ the\ data} = \frac{\#\ of\ detections\ where\ tracking\ error < error\ threshold}{Total\ \#\ of\ Faces\ in\ the\ data}$$



First Pass:
Precision = 77.42%
Recall = 61.61%



Second Pass:
Precision = 72.11%
Recall = 80.27%

- Mean Tracking Error per frame = mean (pixel distance between the 7 annotated points and the corresponding tracked points, ignore the rest).
- Mean Normalized Tracking Error = Mean Tracking Error / Intraocular Distance

Success Criteria: Detection score > -0.30, normalized tracking error < 0.15

Summary of Face Detection and Tracking Performance

Face Detection Performance

Dataset	Approach	Success Rate	Median Score	Precision	Recall
HPV hi-res	**First Pass**	79.34%	0.38	99.26%	79.58%
	Second Pass	95.66%	1.45	**96.54%**	**96.06%**
SHRP2 lo-res	**1X First Pass**	67.22%	0.07	99.64%	64.19%
	1X Second Pass	97.99%	1.36	**98.24%**	**98.59%**
	2X First Pass	79.52%	0.37	99.14%	77.46%
	2X Second Pass	93.49%	1.17	**98.82%**	**92.47%**

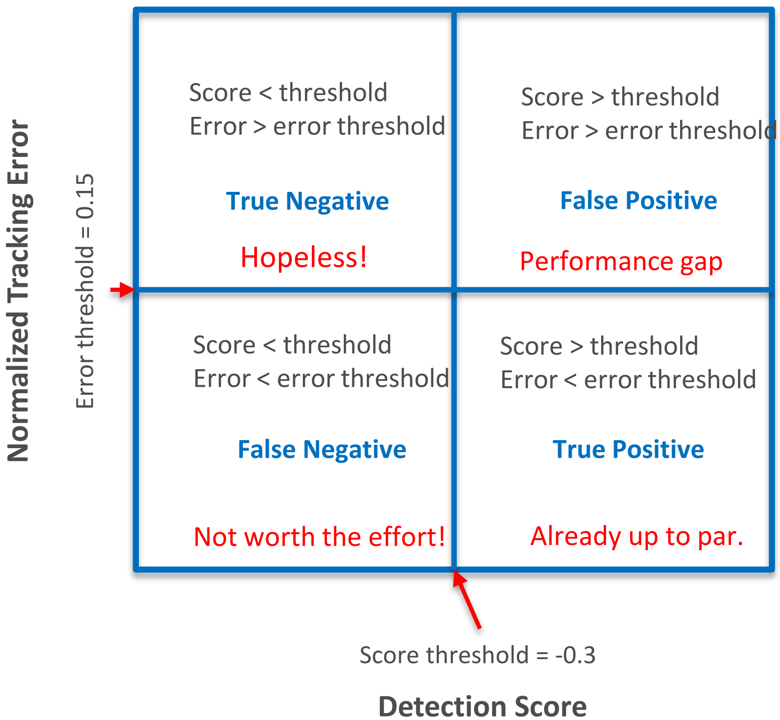
Face Tracking Performance

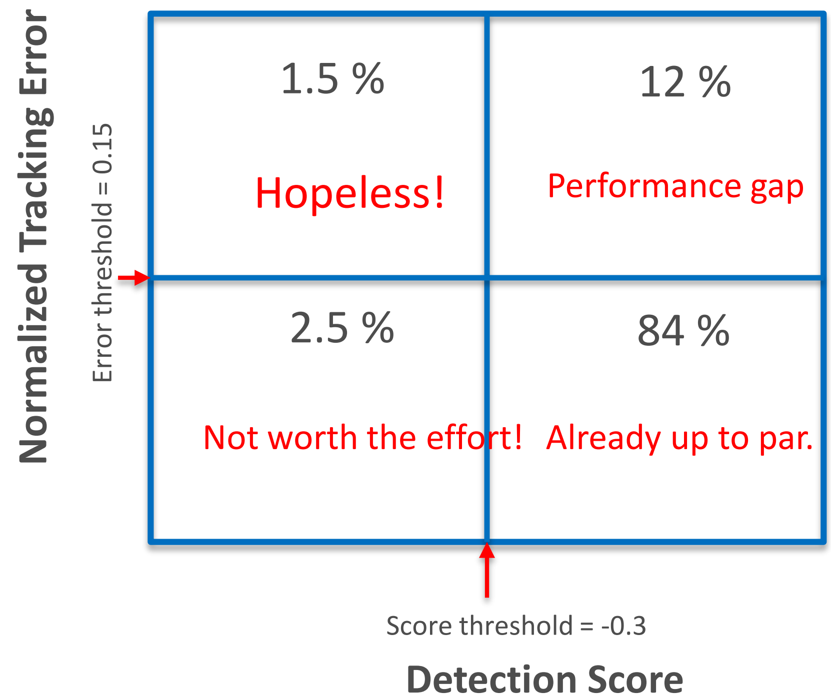
We are able to track the facial features in the SHRP2 lo-res videos fairly well but we are still about 10% below the performance of the hi-res videos (HPV).

Dataset	Approach	Precision	Recall
HPV hi-res	**First Pass**	77.4%	61.6%
	Second Pass	72.1%	**80.3%**
SHRP2 HPV lo-res	**1X First Pass**	51.3%	32.9%
	1X Second Pass	39.2%	38.6%
	2X First Pass	65.4%	49.1%
	2X Second Pass	69.1%	71.6%

- HPV hi-res : 720 x 480
- SHRP2 (HPV) lo-res: 1X = 356 x 240, 2X = 712 x 480

Performance Analysis Quad Chart : End of Program

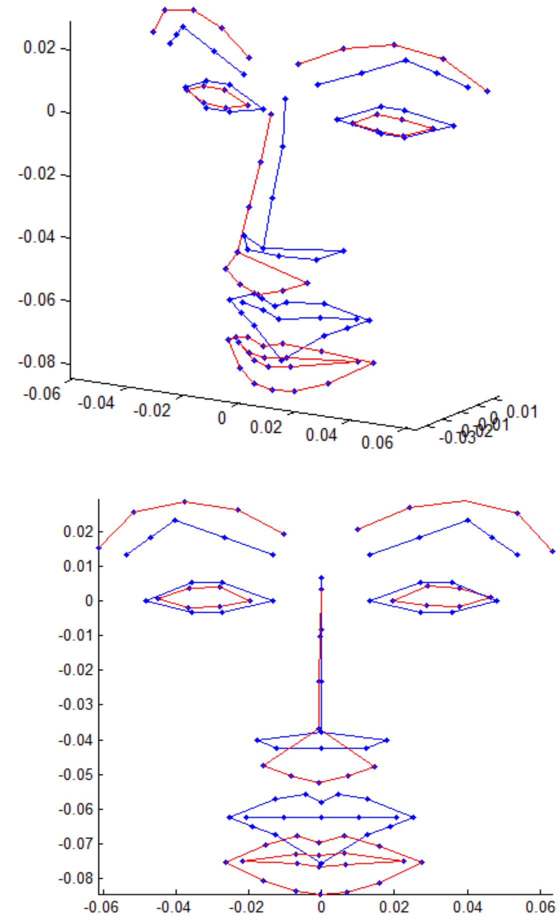




Core Feature: Head/Face Pose Tracking Customizing the Face Model

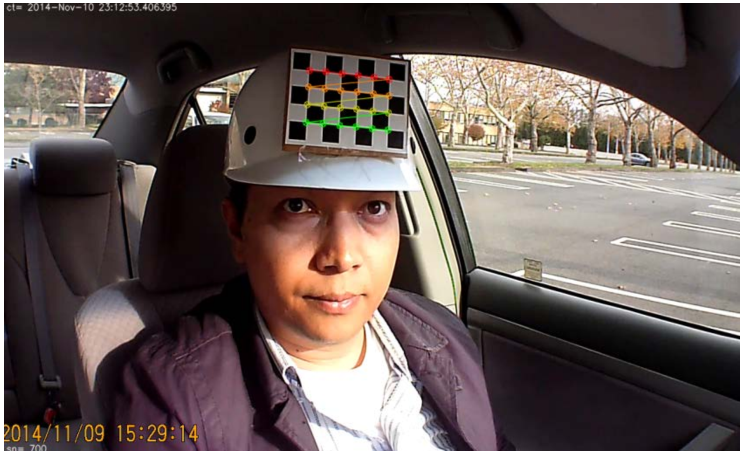
- Reconstruct Face Model from different views of the driver

Collection of tracked landmarks in different poses

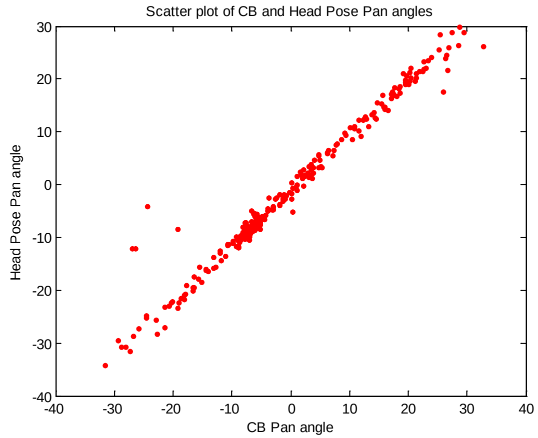


[B] Original average model
[R] Customized face model

Evaluation of Head Pose Accuracies



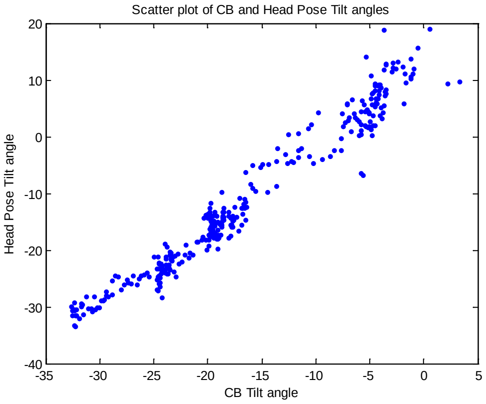
Pan



Errors:

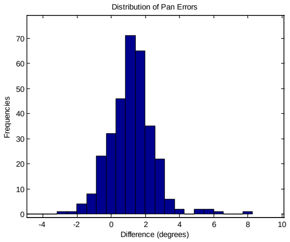
Pan:
std = 2.24 deg

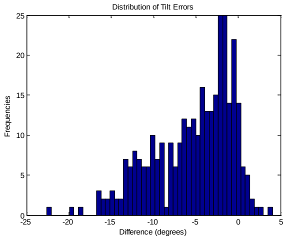
Tilt:
std = 4.70 deg



Tilt

Error Distribution

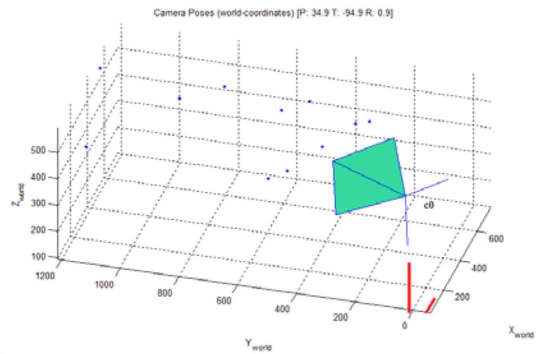




Error Distribution

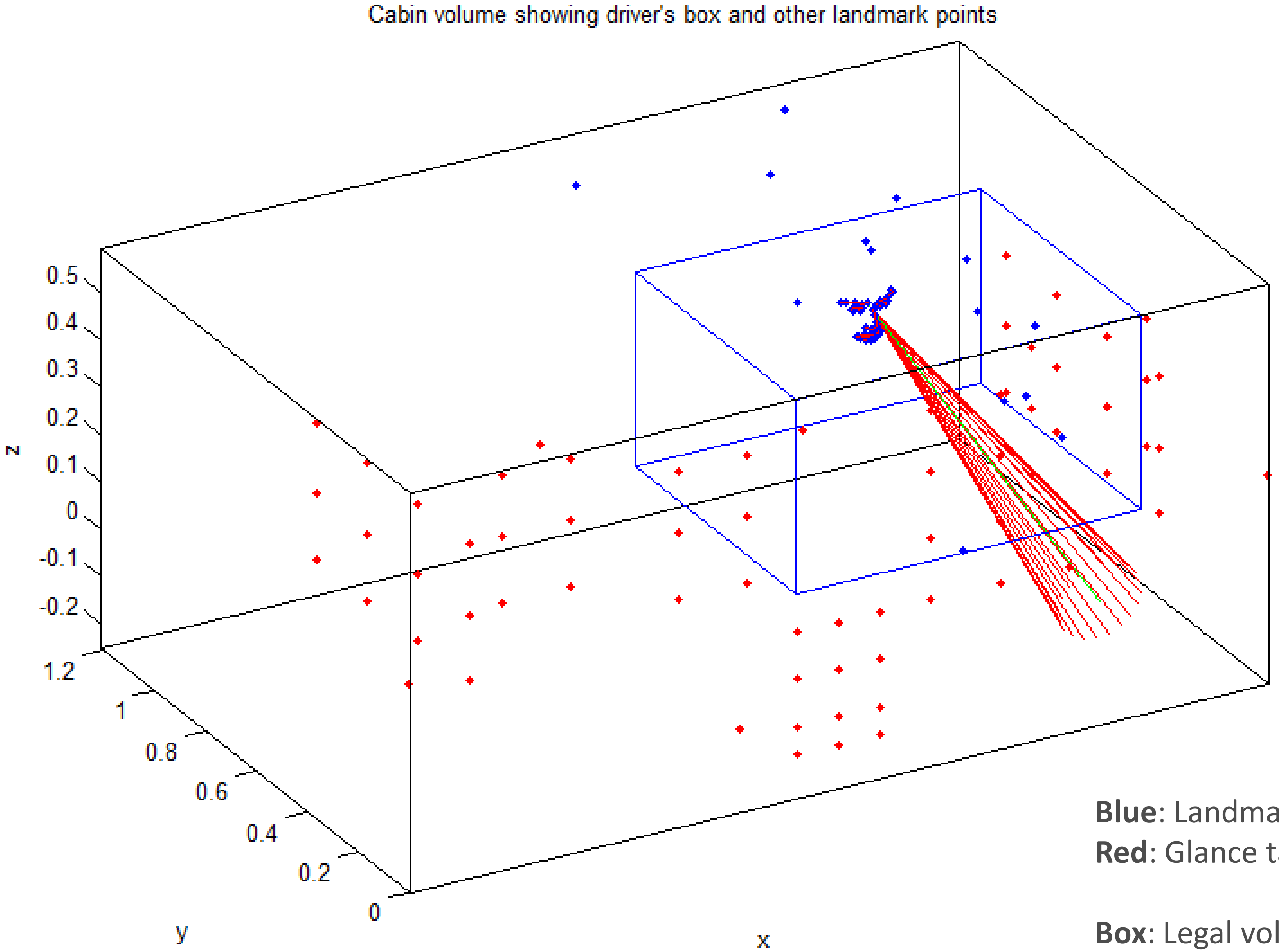
Driver View Camera Extrinsic Calibration Approach

- Calibrate camera relative to the vehicle cabin coordinates
  - Needed for glance target tracking and other estimations that require better geometry
- Kinect and Laser scans available from vehicle interior
  - Laser scans go further into the cabin but point were hard to read
  - Kinect scans were easier to work with but were not extensive
- Vehicle used in the HPV study is the Saab Model
- Camera Intrinsic Matrix is known (from ORNL)



Recognizable Features from the Video

Final Coded Feature: Using Head/Face Pose to Compute 3D Glance Target Vectors (Gaze Monitoring)

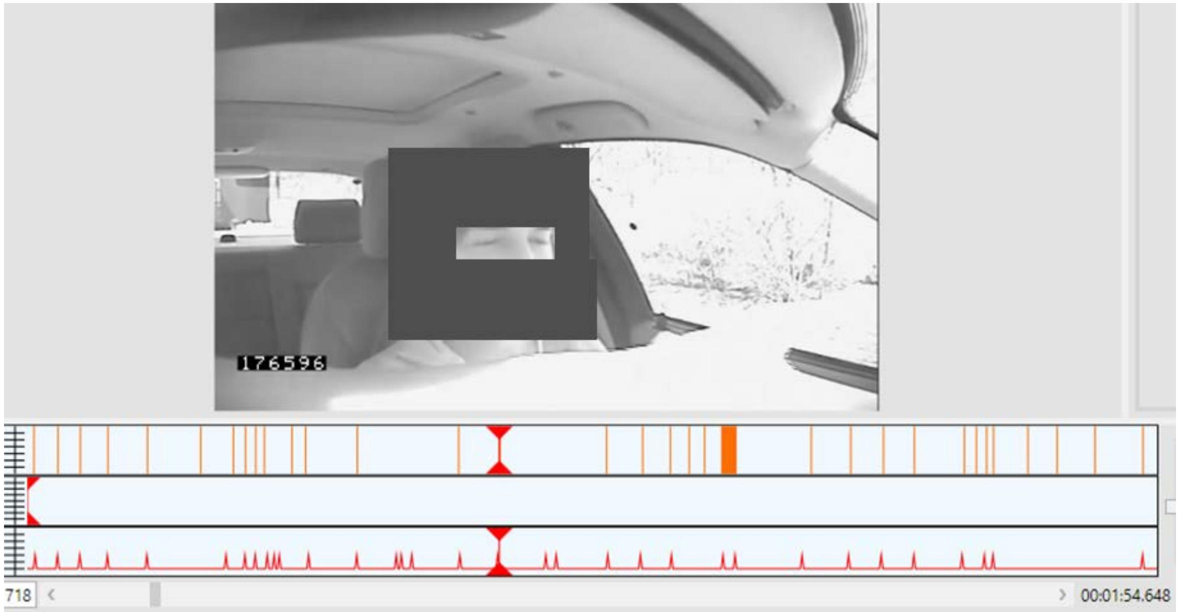


Blue: Landmark points
Red: Glance target points

Box: Legal volume for the driver's head

Intermediate Feature: Eye Blink Monitoring

- Eye Blink Detection
  - Currently based solely on the tracked landmark features
- Used for Blink-Rate Estimation, percentage eyes closed, eye close durations, etc.

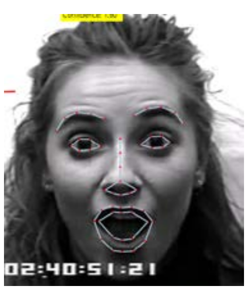


Intermediate Feature: Facial Expression Analysis

- Goal:
  - Try to identify driver anxiety (nervous driving), anger (road rage), etc.

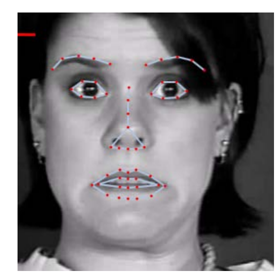
- Seven standard facial expression classes were trained using the Cohn-Kanade+ dataset
  - Neutral, Angry, Contempt, Disgust, Fear, Happy, Sadness, Surprise

- Qualitatively, the only expression that seems to arise in this data is “happy” when the drivers are chatting with the person in the passenger’s seat.



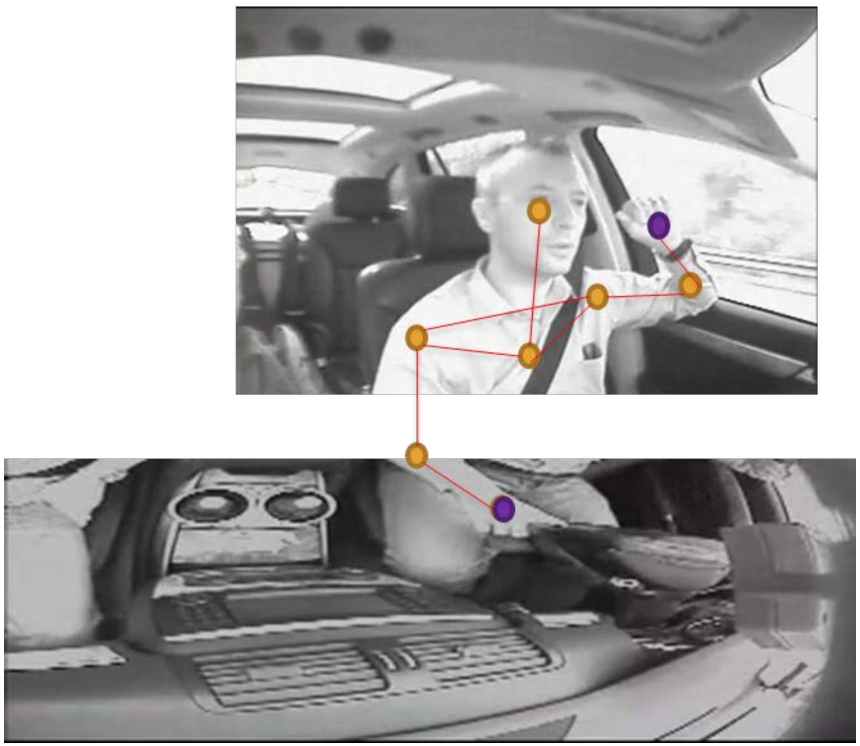


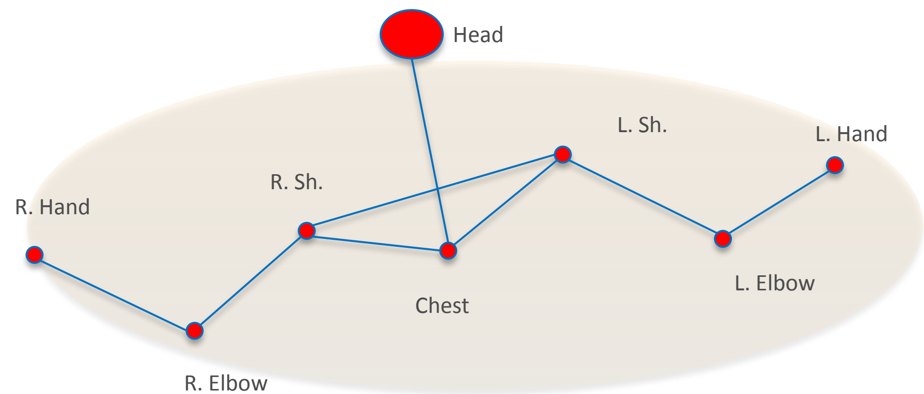




Core Features : Driver's Hands and Upper Body Pose Tracking

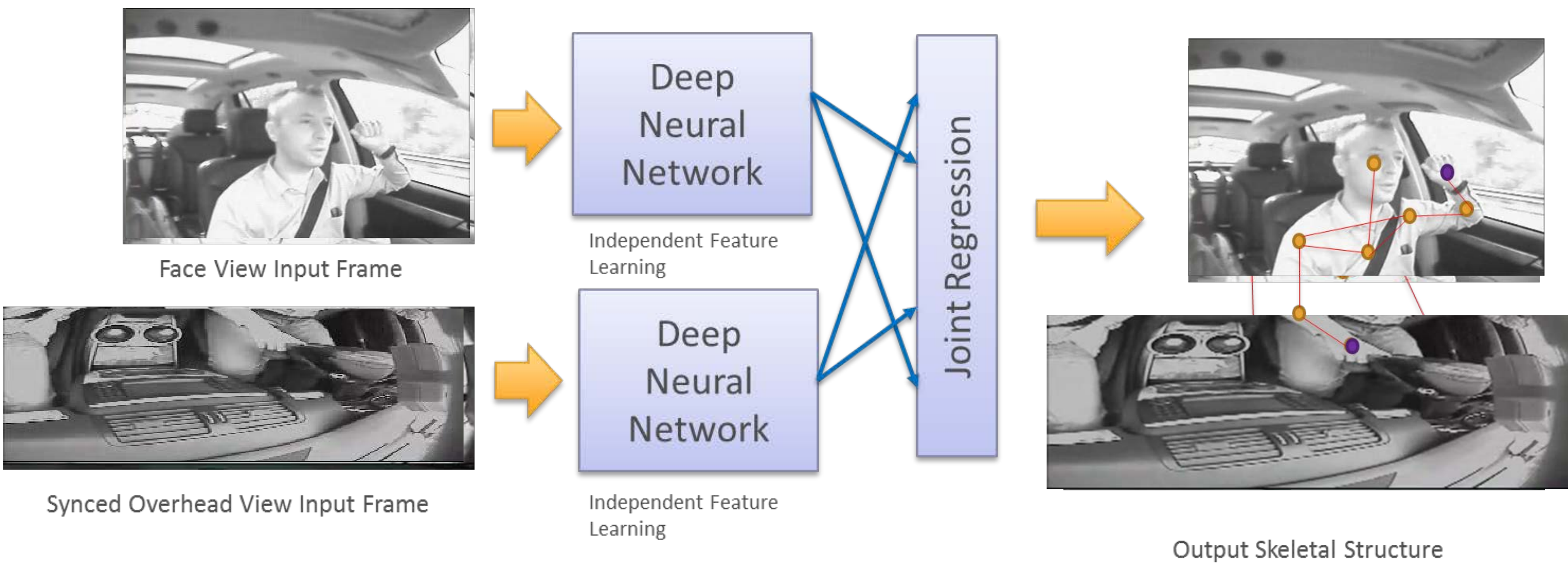
- Goal:
  - Track upper body joints skeleton (to ultimately track driver activity)
  - Jointly from the frontal face view video and the overhead hands view video of the SHRP2 dataset



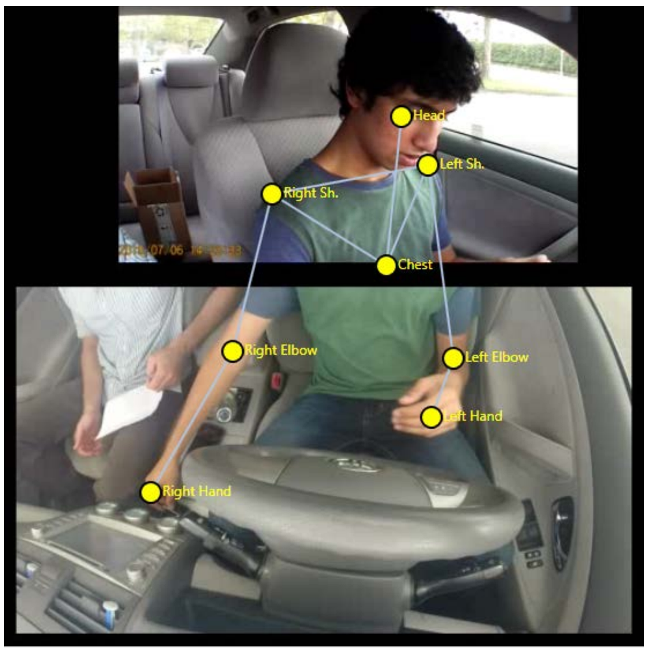


Our Skeletal Representation
for
Upper Body Pose Tracking

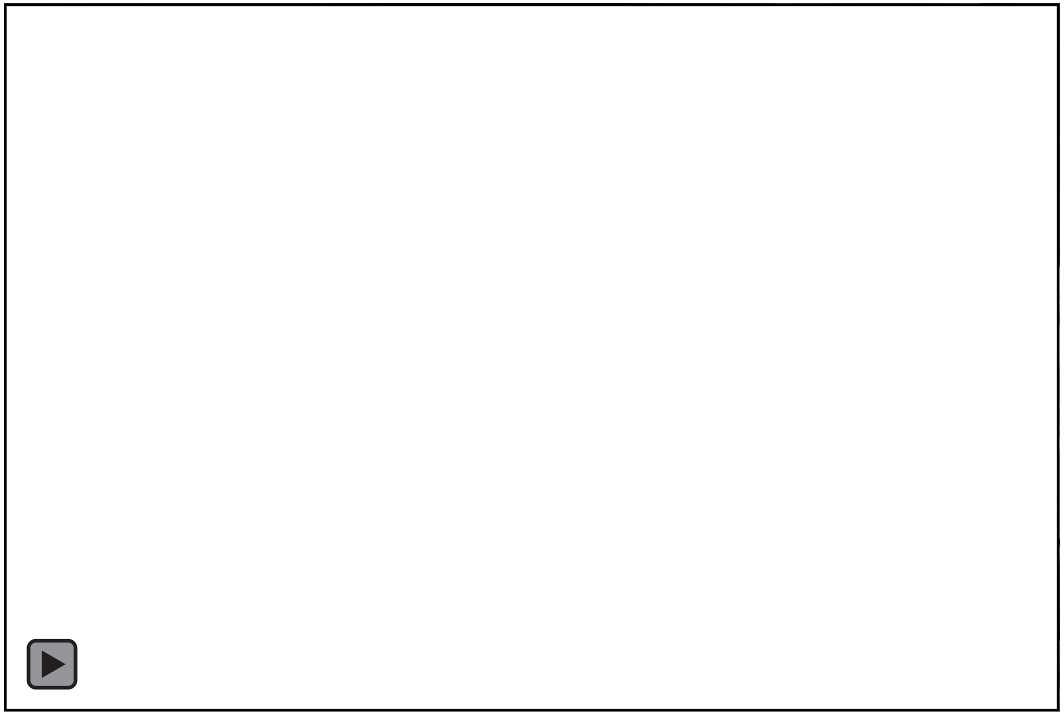
Technical Approach: Deep Pose Algorithm



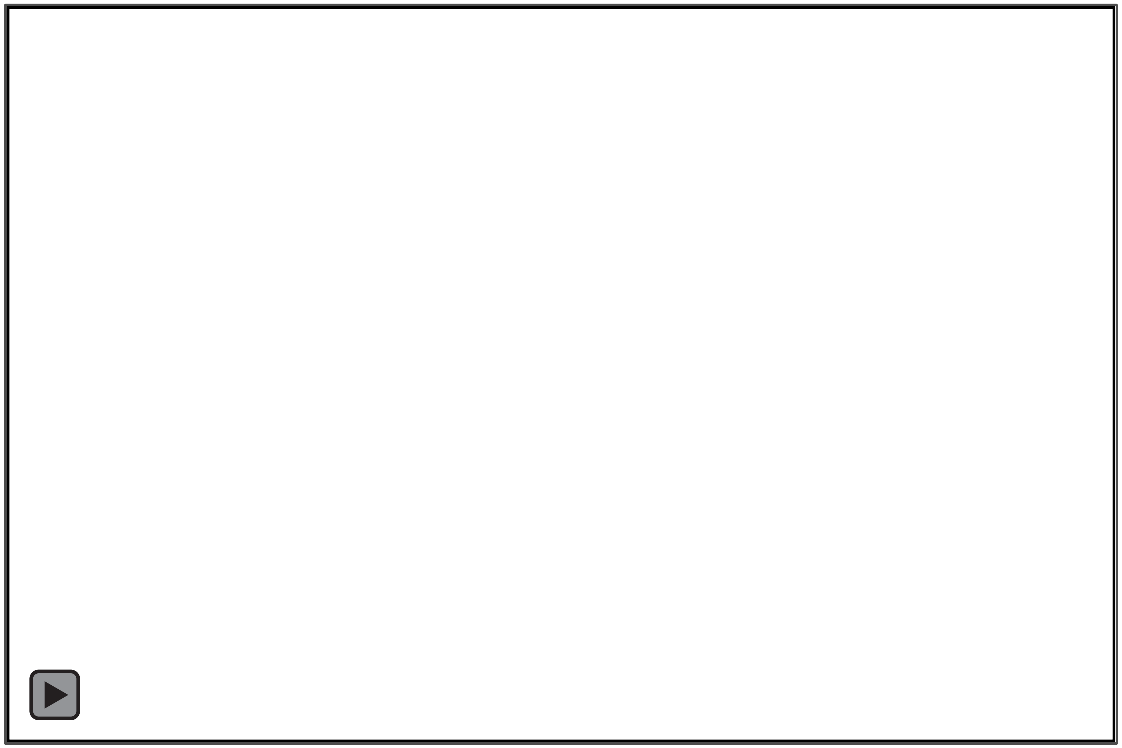
Upper Body Pose Tracking Examples



Local Driving Data



Identity Masked (i.e., codings only) visualization of one example SHRP2 video)



This video shows a visualization of the driver video using only the low-level body tracking information

- facial landmarks, head pose and upper body pose skeleton.

Generating Identity Masked Videos From the Tracked Data (DMask)

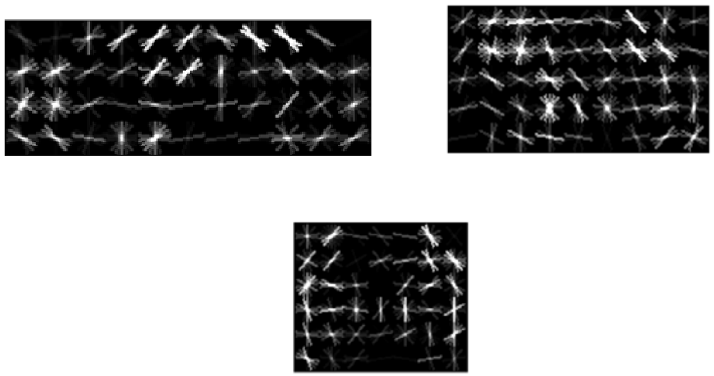
This video shows the motion-transferred virtual avatar rendered over the original video.

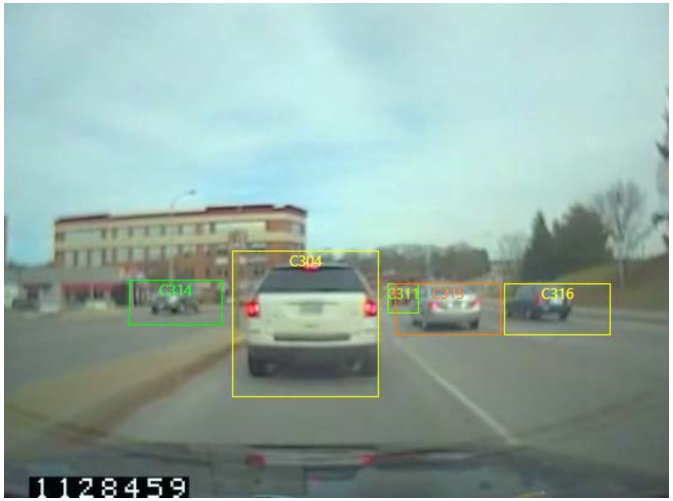
Intermediate Features: Driver Gesture/Action Recognition

(Overall accuracy: 79.83%)

Class	True positive	True Positive + False Positive	True positive + miss detection	Recall	Precision
Make phone call	35	56	42	(83.33%)	(62.5%)
Put on glasses	25	28	29	(86.21%)	(89.29%)
Driving (default)	24	(35	29	(82.76%)	(68.57%)
Adjust mirror	10	12	14	(71.43%)	(83.33%)
Talk to passenger	37	44	44	(84.09%)	(84.09%)
Drink from a cup	24	26	33	(72.73%)	(92.31%)
Rest arm on window	18	20	23	(78.26%)	(90%)
Put on safety belt	25	27	29	(86.21%)	(92.59%)
Take off safety belt	23	32	28	(82.14%)	(71.88%)
Look back – backing up	36	38	41	(87.80%)	(94.74%)
Touch face	24	34	40	(60%)	(70.59%)

Core Contextual Feature: Vehicle Detection and Tracking





Core Contextual Feature: Vehicle Detection and Tracking

Intermediate Feature: Brake Lights/Turn Signal Detection

DCode End Product:
Screenshot of Our DCode Visualization Software Showing Various Automatically Extracted Codings

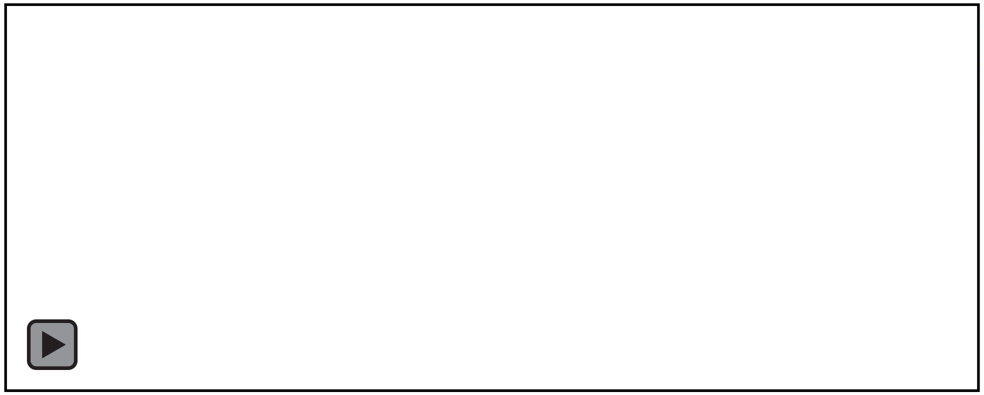
Lessons Learnt and Recommendations: A Computer Vision Perspective

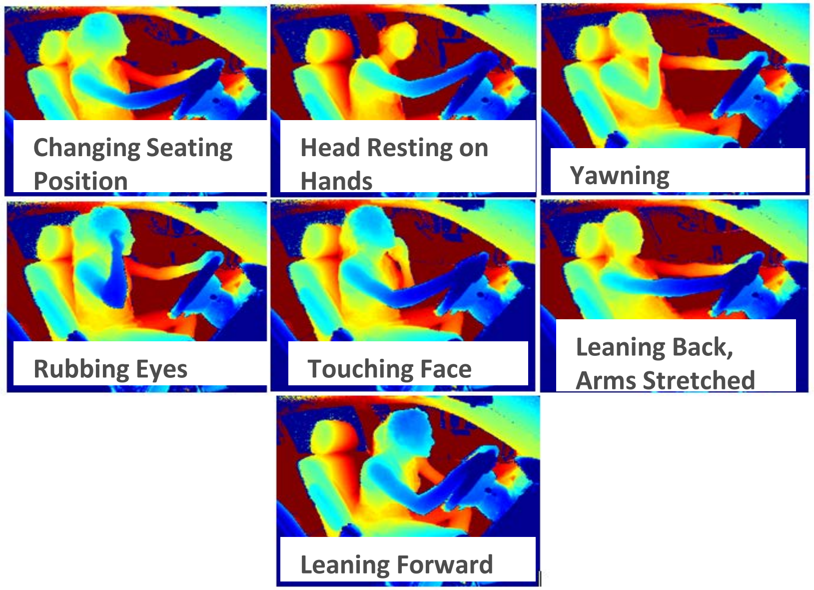
- Video resolution:
  - Tracking performance is a function of resolution up to a point, beyond which the return starts to diminish
  - Resolution vs. FOV: (at least 400x400 pixels on the face)
- Camera position has an impact on the accuracy of tracking
  - Rear view mirror vs steering column vs A-pillar
  - Bottom-up view is better for eye (gaze) tracking (instrument panel, center console, cup holder, cell phones, etc.)
- Illumination management
  - Filter out ambient light as much as possible and use internal illumination
    - Easier to control the quality of the data
  - Helps with managing the glare on glasses.

Lessons Learnt and Recommendations: A Computer Vision Perspective

- Real-time systems (OTS) vs. raw data recording systems (post processing)
  - OTS DMS systems (option to record the metadata only, lower data rates, no legal hassles)
  - Offline data processing allows us to use multi-pass and non-causal data processing approaches, adapt algorithmic parameters for feature extraction (automated coding)

- RGB-d sensors (depth sensing cameras)
  - Allows for more robust upper-body tracking for driver activity monitoring





Thank you.

Headquarters: Silicon Valley

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000

Washington, D.C.

SRI International
1100 Wilson Blvd., Suite 2800
Arlington, VA 22209-3915
703.524.2053

Princeton, New Jersey

SRI International Sarnoff
201 Washington Road
Princeton, NJ 08540
609.734.2553

Additional U.S. and international locations

www.sri.com

FHWA Strategic Highway Research Program -2 (SHRP2)

SHRP2
STRATEGIC HIGHWAY
RESEARCH PROGRAM

- SHRP2 was established by Congress to investigate the underlying causes of highway crashes and congestion in a short–term program of focused research.

- The objective was to identify countermeasures which will significantly improve highway safety through an understanding of driving behaviors.

- Naturalistic Driving Study (NDS) under the SHRP2 program
  - Collected normal driving behavior data
    - 3,400+ drivers
    - 5,400,000+ Trip
    - ~1 Million hours of video data + other metadata

  - There is way too much data for manual coding!
    - FHWA EAR 2A program was created to help develop technologies for automated coding.

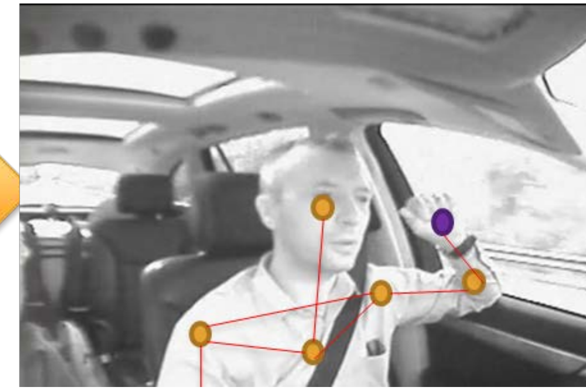
Current Implementation:

Face View Input Frame

Deep Neural Network

Independent Feature Learning

Regression



Overhead View Input Frame

Deep Neural Network

Independent Feature Learning

Regression

Output Skeletal Structure

Overhead View Turned out to be too low quality!

Intermediate Contextual Feature: Brake Lights/Turn Signal Detection