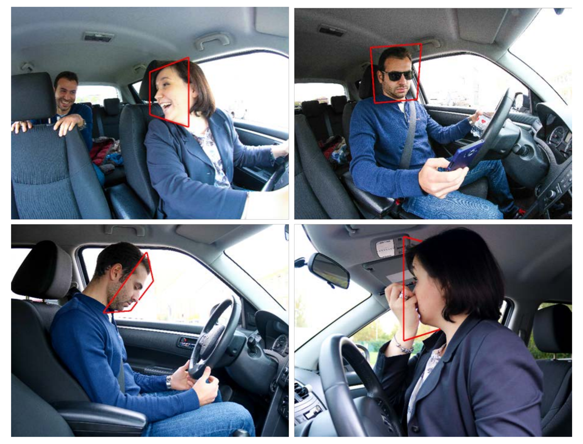
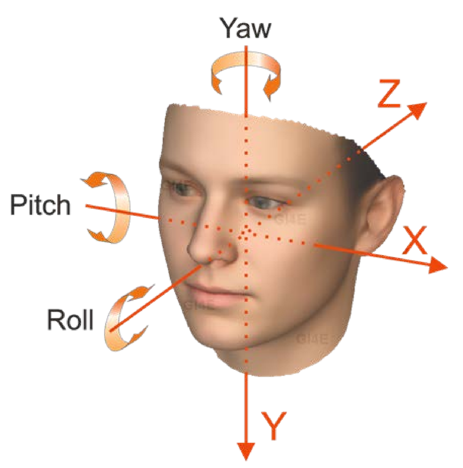
POSEidon: Face-from-Depth for Driver Pose Estimation

Guido Borghi, Marco Venturelli, Roberto Vezzani, Rita Cucchiara
Imagelab - Dipartimento di Ingegneria "Enzo Ferrari " - University of Modena and Reggio Emilia - Italy

Motivations

- We aim at monitoring the driver attention, day and night
- Continuous head pose estimation provides useful cues
- Requirements:
  - Non-Invasive (no wearable devices)
  - High computation speed is mandatory (real-time processing)
  - Independency from external Illumination
  - Embedded systems portability

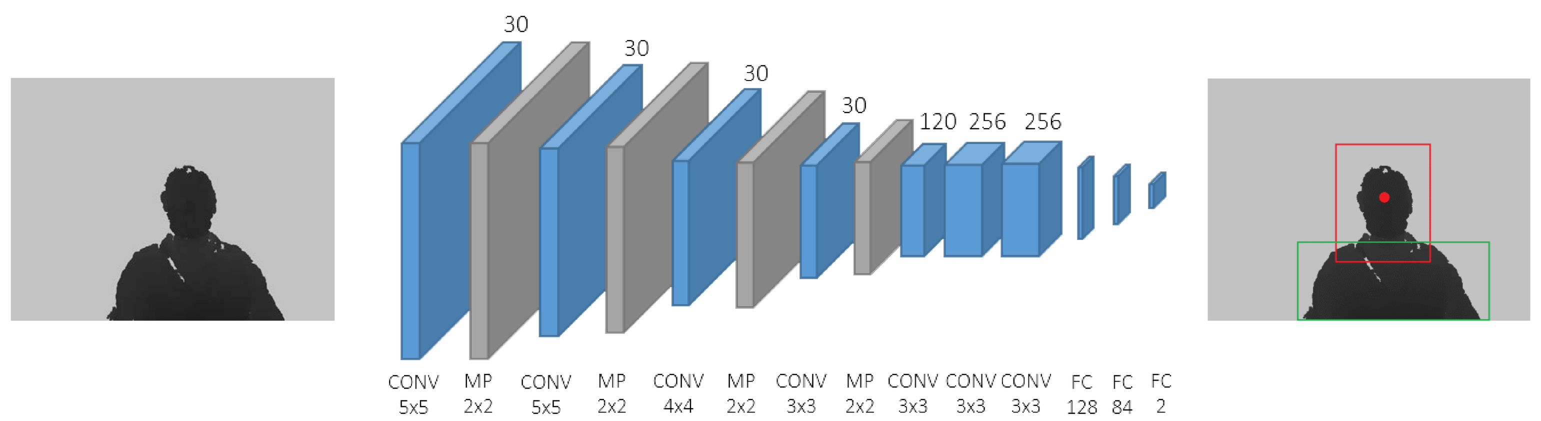




Proposed solution: computer vision based head pose estimation with shallow deep networks on depth images

Head Localization

- **Input**: depth frames
- **Output**: head center position (coordinates *x,y*)



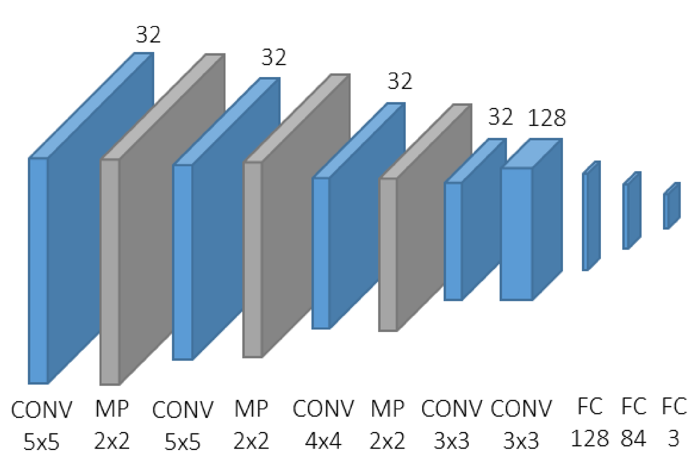
The head size in pixels is estimated given the head center position and the depth (i..e., distance) values around it

Shoulder Pose Estimation

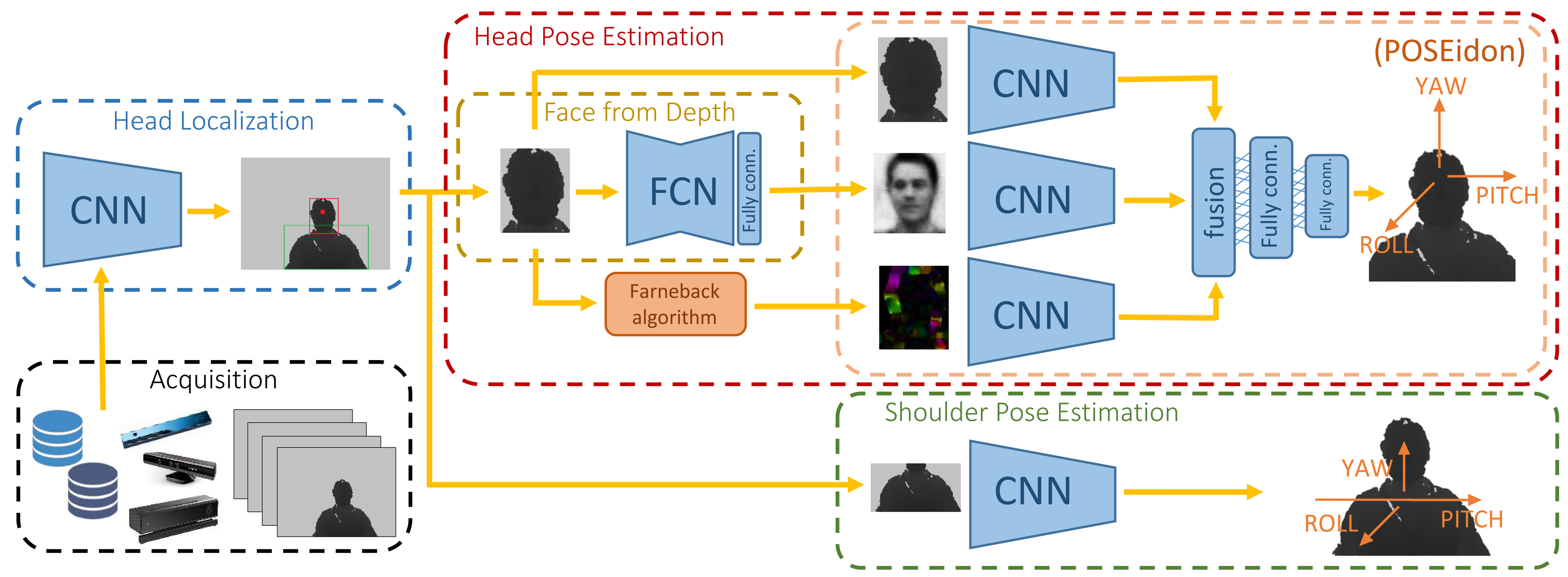
- **Input**: depth frames
- **Output**: 3D shoulder pose angles (*yaw, pitch* and *roll*)

A single network, with the same architecture of CNNs exploited for head pose estimation task

Combined with the head, shoulder pose helps to detect distractions



Overall architecture of the system

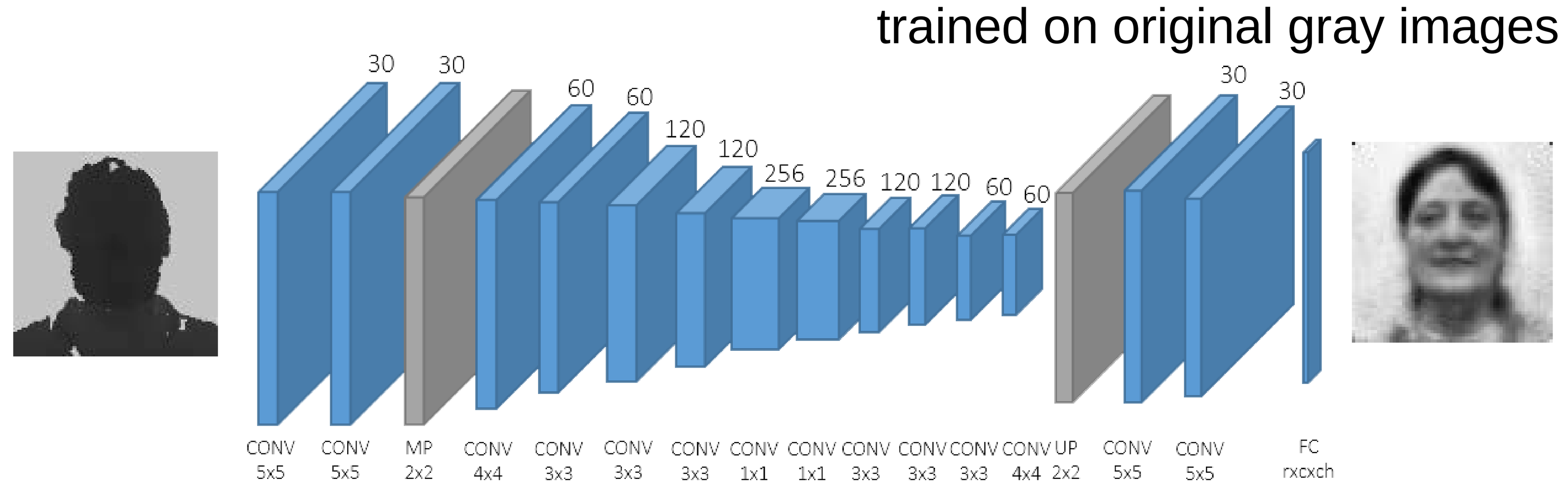


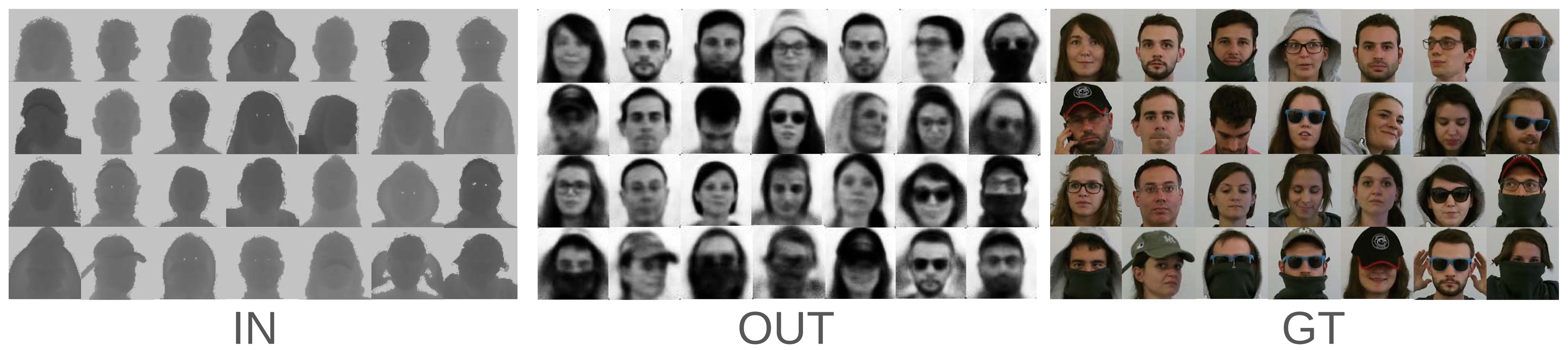
Face-from-Depth

- **Input**: depth frames
- **Output**: gray-level frames

Face-from-Depth generates gray-level images starting from depth.

Output images can be fed to classifiers trained on original gray images

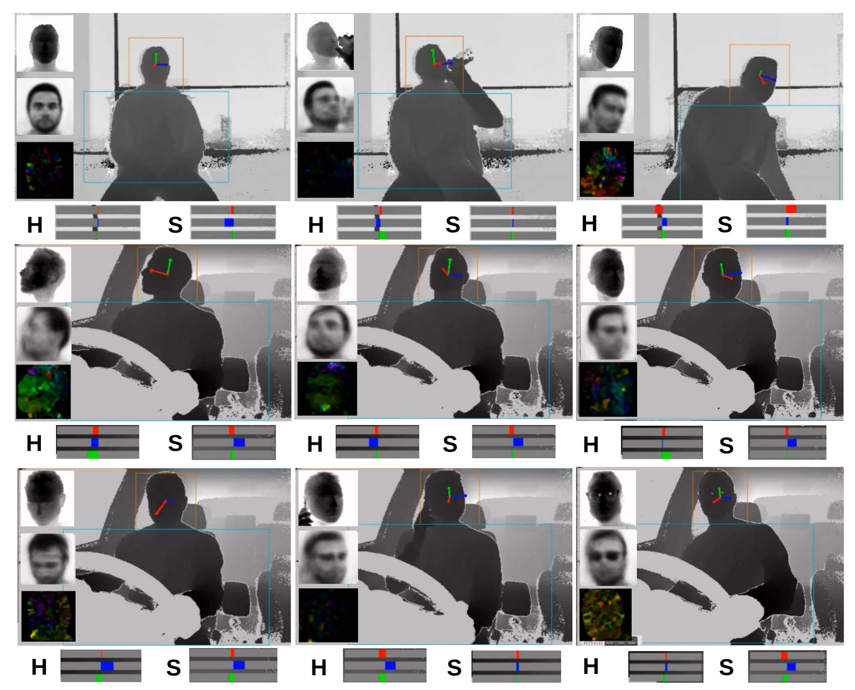




Head Pose Estimation

- **Input**: depth frames, Farneback Optical Flow images, Face-from-Depth images
- **Output**: 3D head pose angles (*yaw, pitch* and *roll*)

The overall POSEidon network is obtained as a fusion of 3 CNNs, individually trained for a regression on the 3D pose angles. Three additional fully connected layers are used to merge the contributions



Experimental results

1. **public datasets** exploited:
   - **Biwi** Kinect Head Pose: 15k images
   - **ICT-3DHP** database: 10k images
2. Pandora dataset
   - **Annotation of shoulder angles**
   - **Wide** angle ranges
   - Challenging **camouflage and postures**
   - **Deep learning** oriented (250k images)
   - High quality **ToF** data (Kinect v2)

Pandora website

Sample frames from Pandora dataset

HEAD POSE ESTIMATION ERROR [EULER ANGLES]

Architecture	Input	Cropping	Fusion	Head Pitch	Head Roll	Head Yaw	Accuracy
Single CNN	depth		-	8.1 ± 7.1	6.2 ± 6.3	11.7 ± 12.2	0.553
	depth	√	-	6.5 ± 6.6	5.4 ± 5.1	10.4 ± 11.8	0.646
	FfD	√	-	6.8 ± 7.0	5.7 ± 5.7	10.5 ± 14.6	0.647
	gray-level	√	-	7.1 ± 6.6	5.6 ± 5.8	9.0 ± 10.9	0.639
	MI	√	-	7.7 ± 7.5	5.3 ± 5.7	10.0 ± 12.5	0.609
Double CNN	depth + FfD	√	concat	5.6 ± 5.0	4.9 ± 5.0	9.8 ± 13.4	0.698
	depth + MI	√	concat	6.0 ± 6.1	4.5 ± 4.8	9.2 ±11.5	0.690
POSEidon	depth + FfD + MI	√	concat	6.3 ± 6.1	5.0 ± 5.0	10.6 ±14.2	0.657
	depth + FfD + MI	√	mul+concat	5.6 ± 5.6	4.9 ± 5.2	9.1± 11.9	0.712
	depth + FfD + MI	√	conv+concat	5.7 ± 5.6	4.9 ± 5.1	9.0 ± 11.9	0.715

Results on *Pandora*

HEAD POSE ESTIMATION ERROR [EULER ANGLES]

Method	Year	Data	Pitch	Roll	Yaw	Avg
Fanelli *et al.*	2011	Depth	8.5 ± 9.9	7.9 ± 8.3	8.9 ± 13.0	8.43 ± 10.4
Yang *et al.*	2012	RGB + Depth	9.1 ± 7.4	7.4 ± 4.9	8.9 ± 8.3	8.5 ± 6.9
Padeleris *et al.*	2012	Depth	6.6	6.7	11.1	8.1
Rekik *et al.*	2013	RGB + Depth	4.3	5.2	5.1	4.9
Baltrusaitis *et al.*	2012	RGB + Depth	5.1	11.3	6.3	7.6
Ahn *et al.*	2014	RGB	3.4 ± 2.9	2.6 ± 2.5	2.8 ± 2.4	2.9 ± 2.6
Martin *et al.*	2014	Depth	2.5	2.6	3.6	2.9
Saeed *et al.*	2015	RGB + Depth	5.0 ± 5.8	4.3 ± 4.6	3.9 ± 4.2	4.4 ± 4.9
Papazov *et al.*	2015	Depth	2.5 ± 7.4	3.8 ± 16.0	3.0 ± 9.6	4.0 ± 11.0
Drouard *et al.*	2015	RGB	5.9 ± 4.8	4.7 ± 4.6	4.9 ± 4.1	5.2 ± 4.5
Meyer *et al.*	2015	Depth	2.4	2.1	2.1	2.2
Liu *et al.*	2016	RGB	6.0 ± 5.8	5.7 ± 7.3	6.1 ± 5.2	5.9 ± 6.1
POSEidon	2016	Depth	**1.6 ± 1.7**	**1.8 ± 1.8**	**1.7 ± 1.5**	**1.7 ± 1.7**

Results on *Biwi*

Parameters R_x	Parameters R_y	Shoulders Pitch	Shoulders Roll	Shoulders Yaw	Accuracy
No crop		2.5 ± 2.3	3.0 ±2.6	3.7 ± 3.4	0.877
700	250	2.9 ± 2.6	2.6 ±2.5	4.0 ± 4.0	0.845
850	250	2.4 ± 2.2	2.5 ±2.2	3.1 ± 3.1	0.911
850	500	**2.2 ± 2.1**	**2.3 ±2.1**	**2.9 ± 2.9**	**0.924**

Shoulder pose estimation on *Pandora*

The framework works at 30 fps on a desktop with GPU, while it processes around 10 fps on embedded devices.

ACKNOWLEGMENTS - This work has been carried out within the project "FAR2015 - Monitoring the car drivers attention with multisensory systems, computer vision and machine learning" funded by the University of Modena and Reggio Emilia. We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

project website

Video Demo