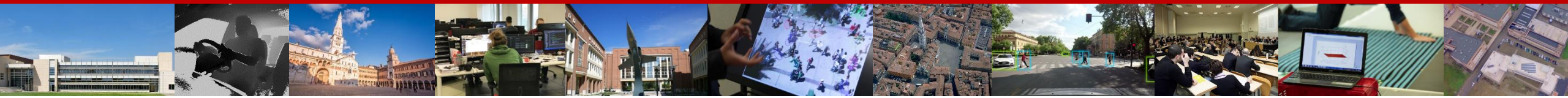
Deep Depth Vision for Driver Attention Monitoring



Guido Borghi

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

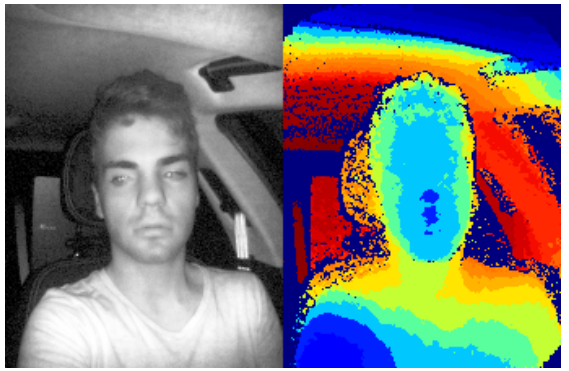
Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

What is a Depth Map?

- A **depth map** is an image, or an image channel, that **contains information about the distance between two objects**, *e.g.* the acquisition device and a surface into the acquired scene, *i.e.* an object visible from the camera's point of view
- From a **2D perspective**, depth maps are usually coded as a **gray-level image**, *i.e.* a single channel-image with a 0 - 255 range. **However, each sensors has its own type.**
- From a **3D perspective**, depth map is **a projection of a point cloud**, in which every point contains the 3D position in respect to the camera coordinate system
- In the literature, depth maps are also referred as *depth images*, *range images* and *2.5D images*

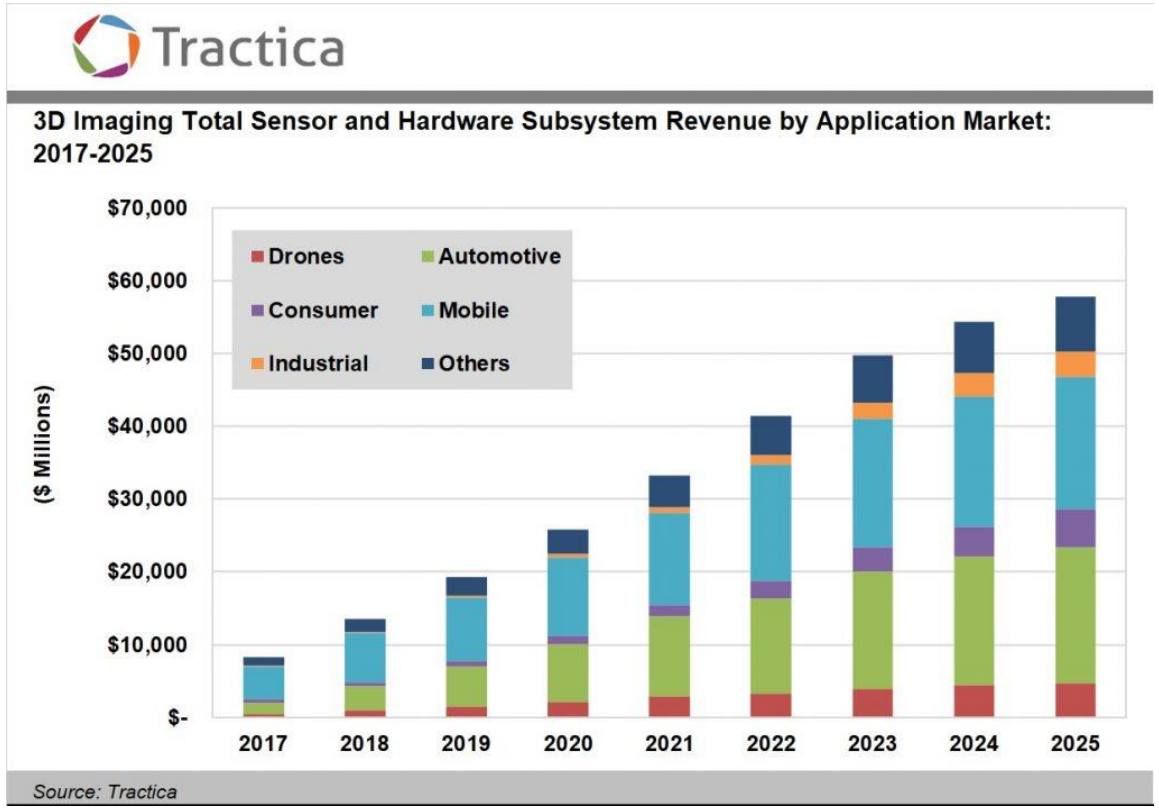
IF Depth Map

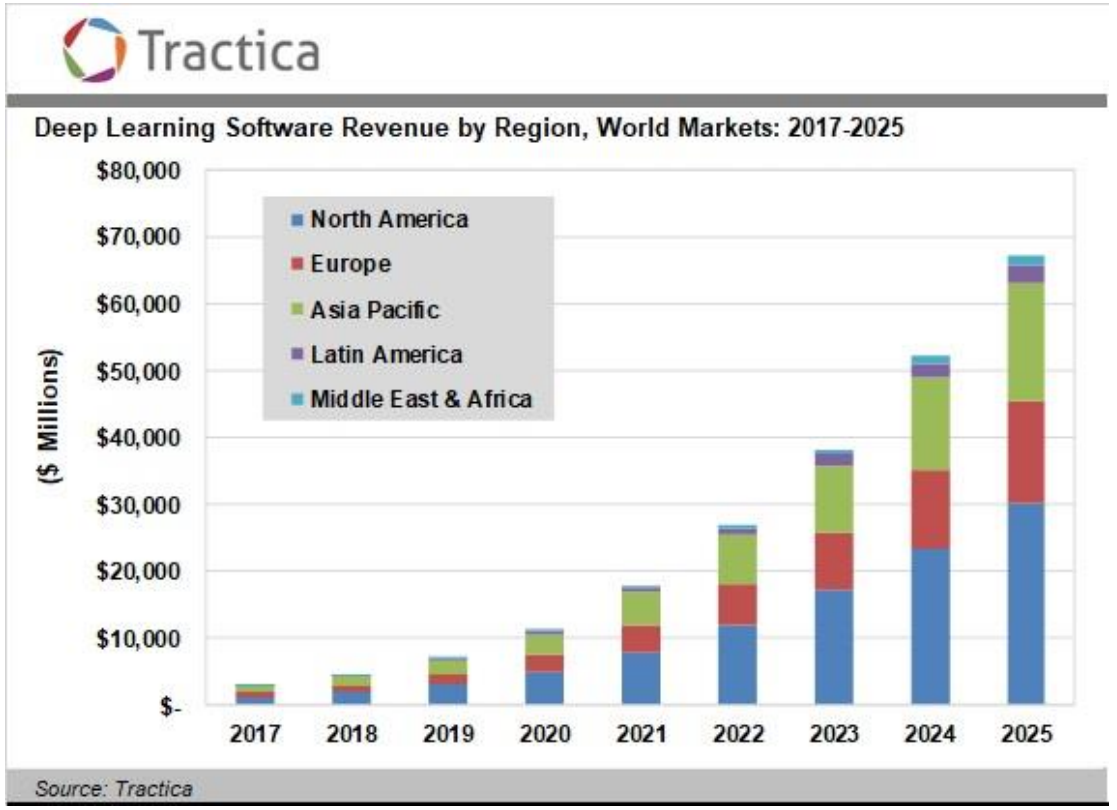
IF Depth Map



Depth Devices

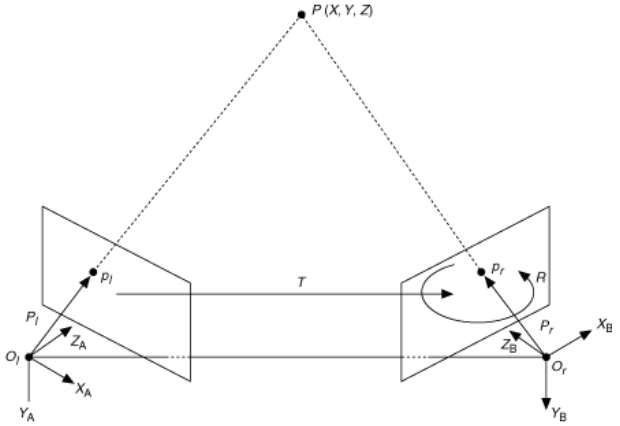
- **Depth Sensors**: devices that are able to provide in output distances
- Recently, a lot of new depth sensors have been introduced in the market
- A new trend is acquiring increasing importance: **Deep Learning** + **Depth Maps**

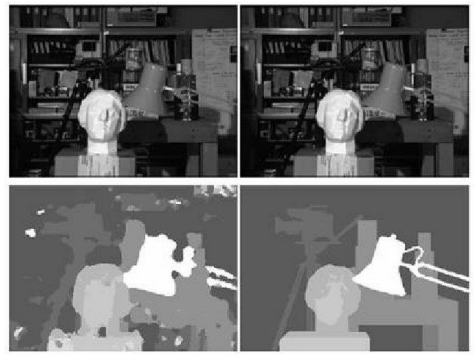




Depth Devices

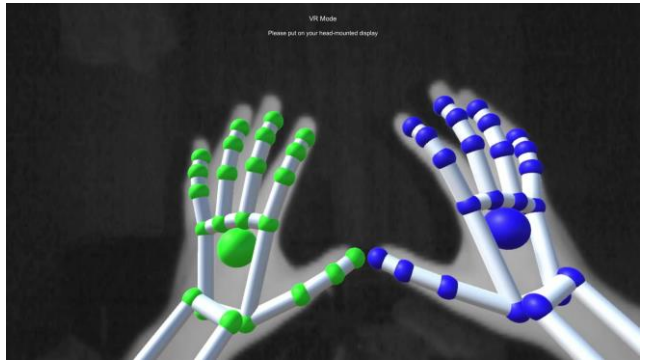
- **Stereo Cameras**:
  - The base concept is similar to what happens in the **human body with the eyes**
  - Two similar cameras are placed in a **fixed distance** on the same plane
  - The **disparity map** of the scene is computed by combining acquired images of these two different intensity (gray-level or RGB) cameras, resolving the so called **correspondence problem**
  - Given a pair of rectified images, it is possible to retrieve the distance of a point in the scene applying a **triangulation method** on a set of corresponding points that lying on *epipolar* lines.





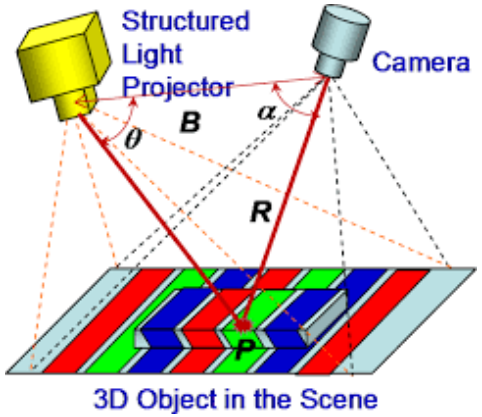
Depth Devices

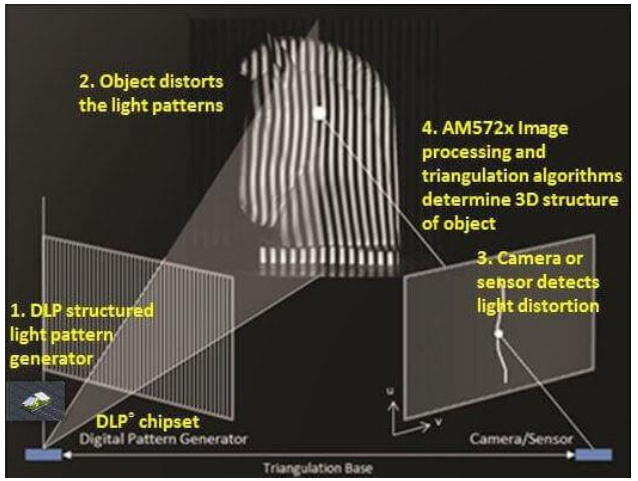
- Example of Stereo Camera: the *Leap Motion* device
  - **2 infrared cameras** with a spatial resolution of **640x240**
  - **Up to 200 frame per second**
  - Field of view: 135° (**fish-eye lens**)
  - **Small Factor Size**: 7 x 1.2 x 3 mm
  - Only **32g** of weight
  - *SDK* for real time hand tracking (robust and accurate)



Depth Devices

- **Structured Light** devices:
  - These scanners project a **specific pattern** inside of the scene
  - The **deformation** in the projected pattern introduced by the objects present inside of the scene is analyzed, through appropriate geometric transformations, **to return for every projected point its 3D position**
  - The hardware of these devices includes a **laser projector** and a **sensor** that is sensitive to the corresponding bandwidth

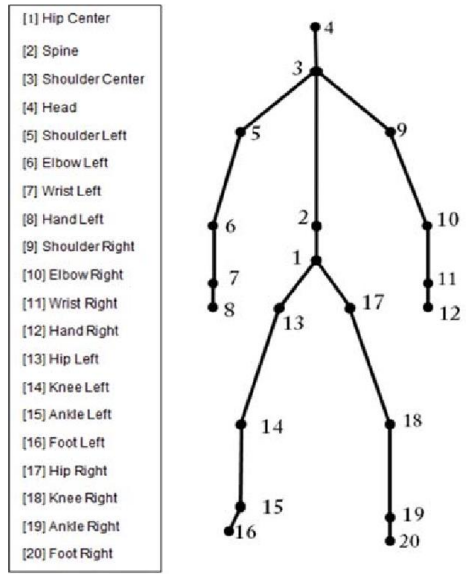


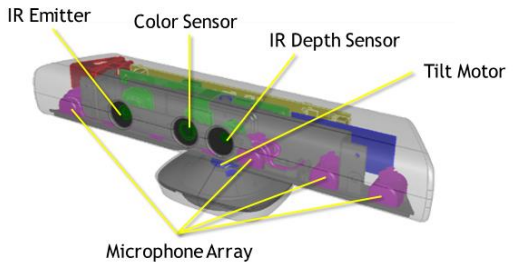


Depth Devices

Example of Structured Light device: the *Microsoft Kinect* (first version)

- **RGB** camera: 640x480 up to 30fps
- CMOS **depth** sensor (320x240)
- **Range**: 0.4 – 4.5/6 m
- **2 full skeleton tracked** (**20 joints**)
- Power consumption: 2.5W
- Field of view: 57° x 43°
- **Tilt motor**

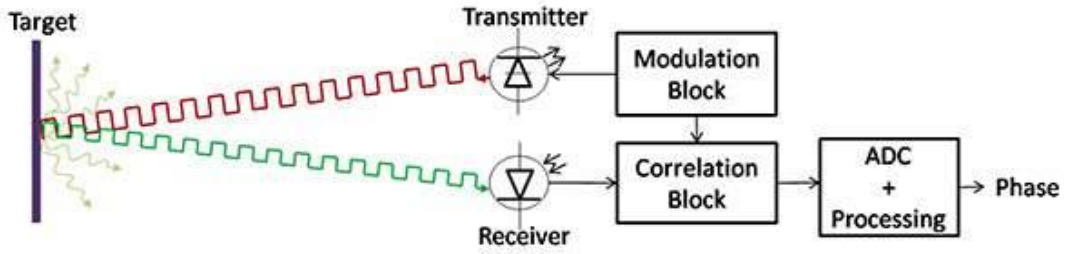




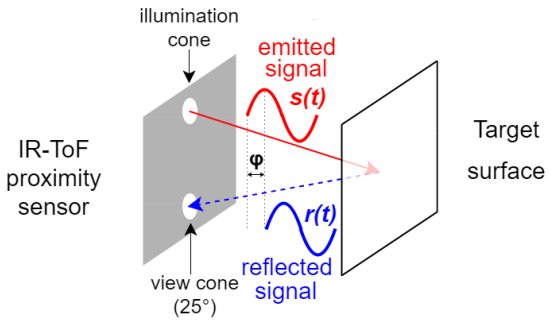
Depth Devices

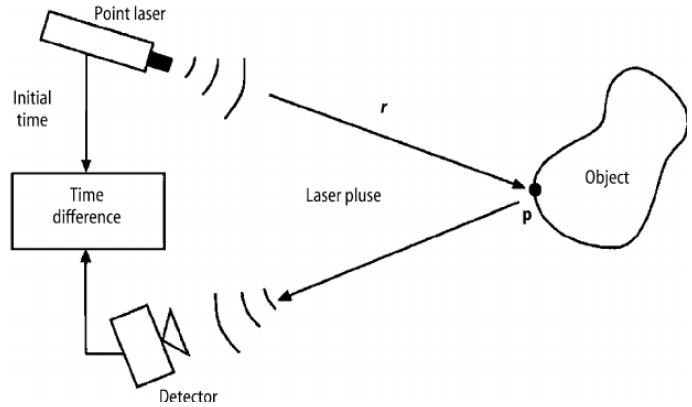
Time-of-Flight devices:

- The distance is computed measuring the **time interval** (the **phase difference**) taken for infrared light to be reflected by the object in the scene



- Also in this case, it is necessary to have a **laser projector** and a **sensor** sensitive to the corresponding bandwidth

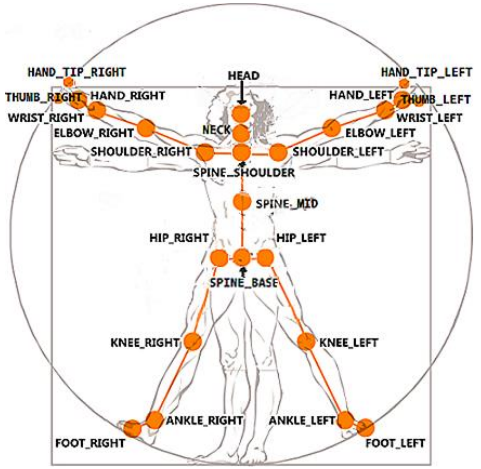




Example of ToF device

Example of ToF device: the *Microsoft Kinect One* (second version):

- **RGB** camera: 1920x1080 up to 30fps
- CMOS **depth** sensor (512x424)
- Range: 0.5 – 5 m
- **6 full skeleton tracked** (**26 joints**)
- Power consumption: 2.5W
- Field of view: 70° x 60°
- **No tilt motor**



Example of ToF

Examples of ToF devices:

- *CamBoard Pico Flexx*
  - **Depth** sensors with a spatial resolution of 224x171
  - Only 68 x 17 x 7.35 mm (**8g**)
  - Up to **45 fps**
  - Range: 0.1 – 4m

- *Pico Zense DCAM710*
  - 69mm x 25mm x 21.5mm
  - **Depth resolution: 640 * 480 @ 30FPS**
  - RGB resolution: 1920 * 1080 @ 30FPS
  - Viewing angle: 69 ° (horizontal) 51 ° (vertical)

1. https://pmdtec.com/picofamily/flexx/
2. https://www.pico-interactive.com/zense

Depth Device Comparison

- Since depth devices are mainly based on the infrared light, depth maps are useful in systems that require:
  - **Light Invariance**: vision-based systems have to be reliable even in presence of light changes
- This is the case of the **automotive** context, in which the light invariance is needed in case of night, tunnels and bad weather conditions.
- The automotive context has some other requirements satisfied by (new) depth devices:
  - **Non-invasiveness**: driver's movements and gaze must not be impeded during the driving activity
  - **Real Time performance**: monitoring and interaction systems have to quickly detect anomalies and provide a fast feedback

Depth and Deep Learning

- Issues using depth maps as input for CNN:
  - **Every sensor has its own depth map**
    - Ad hoc training phases
    - New data to collect
  - Usually they are used **as standard intensity 2D images**
    - Perspective problems
  - **Alternative ways** to use depth maps as input:
    - Voxel
    - HHA (horizontal disparity, height above ground, and the angle the pixel's local surfacenormal makes with the inferred gravity direction)
    - …

1. *"Learning Rich Features from RGB-D Images forObject Detection and Segmentation"* https://arxiv.org/abs/1407.5736

Driver Distraction

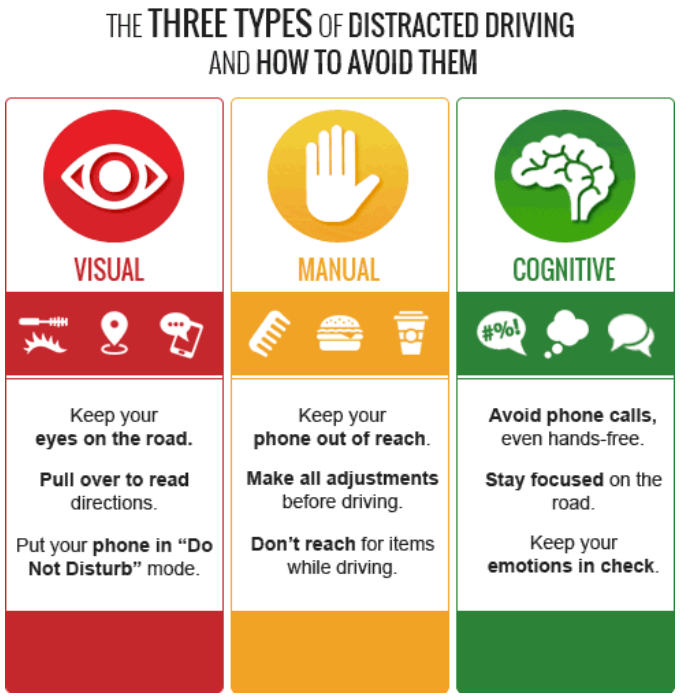
- **What is Driver Distraction**?
  - It is a form of **inattention** (*i.e. failure to pay attention*)
- **Definition**:
  - «*The diversion of attention away from activities critical* ***for safe driving toward a competing activity****, which may result in insufficient or no attention to activities critical for safe driving*»
    (Regan, Hallet & Gordon, 2011)
- Generally, it is **hard** to find a **complete** and accurate description
  - It is also difficult to investigate and find **general solutions** for the problem

Why studying Driver Distraction?

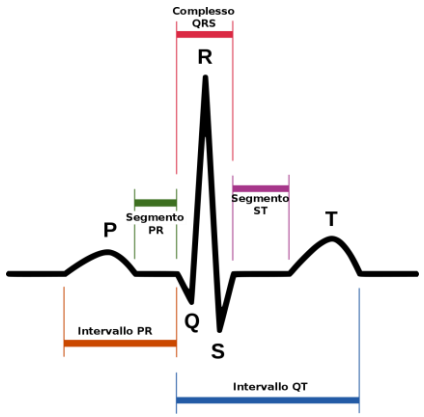
- About **5% to 25%** of **car crashes** have been attributed to driver distraction (and more…)
- Drivers spend about **25-30% of total driving time on distracting activities**
  - 50% concerns **conversation** with a passenger
  - 30% concerns distraction **outside** the vehicle
  - 20% is a **technology-related** type of distraction
- About **70%** (!) of truck crashes are related to driver distraction

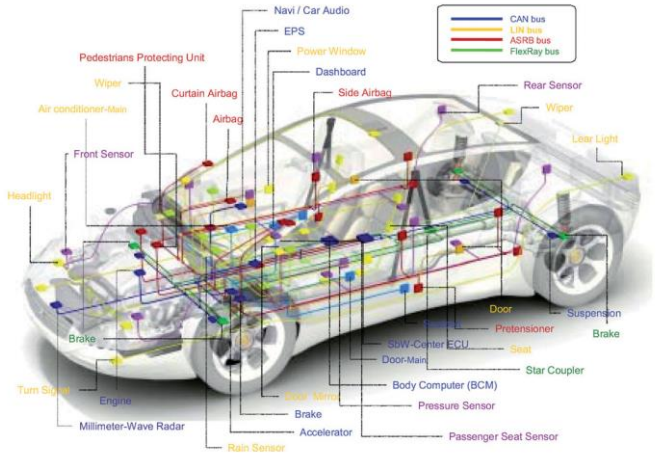
- Generally, there are **four types of distractions**:
  - **Visual Distraction**: driver is looking away from the roadway
  - **Biomechanical Distraction**: driver's hands are not on the steering wheel
  - **Auditory Distraction**: driver is responding to a ringing cell phone
  - **Cognitive Distraction**: driver is not focused on driving activity

- Activities that cause **visual distraction** (*e.g.* looking away from the road during texting) appear to be the **most dangerous**



- Driver Distraction can be monitored through:
  - **Physiological signals**: *electroencephalography, electrocardiography and electromyography* (collected with sensors placed usually inside the steering wheel or the car seat)
  - **Vehicle signals**: parameters acquired from the car bus (*e.g.* steering wheel angles)
  - **Physical signals**: image acquisition and elaboration (our field)

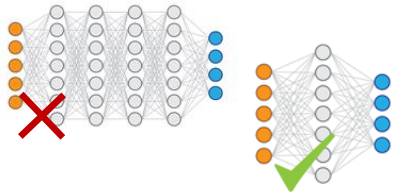




Summarizing

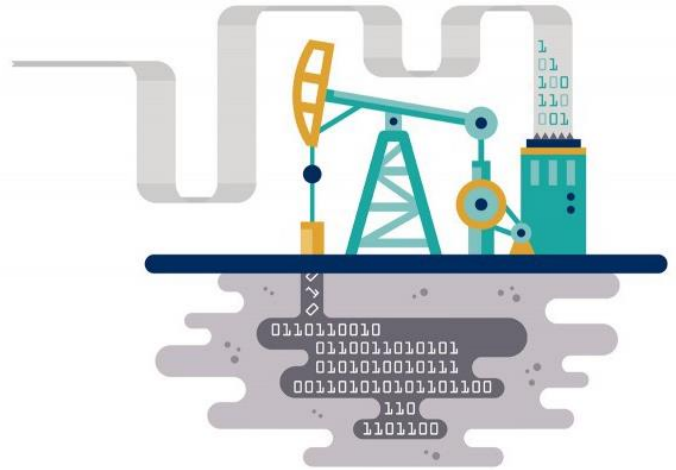
We propose to use Computer Vision and Artificial Intelligence for the Driver Monitoring task

Requirements	Proposed Solutions
Light Invariance Vision-based systems have to be reliable even in presence of *dramatic* light changes	**Depth Maps** Acquired by *IR* sensors. Each pixel contains the value of the distance
Non-invasinevess Driver's movements and gaze *must* not be impeded during the driving activity	**New Sensors** New *cheap* and *accurate* depth sensors with a *small factor form* (*Intel RealSense*, *Pico Flexx*…)
Real-time perfomance Monitoring and interaction systems have to *quickly* detect anomalies and provide feedback	**Design Strategy** *Shallow* Deep Neural Networks Graphics Processing Units



Summary

- Here the tasks related to Driver Monitoring:
  - Face Translation
  - Head Pose Estimation
  - Facial Landmark Localization
  - Head Detection



- Deep learning-based techniques
  - First works in late ‘70, but limited by *mathematical* and *computational* problems
  - Today, it is a *revolution* in the Computer Vision (but not only) field
  - Powerful models, but there are some limitations related to:
    - Availability of a **huge amount** of training data
    - Data *must* be **annotated** (for *supervised* approaches)

    → «Data is the new oil[1]»
    → We collected **new depth-based** datasets

    - **High** computational power needed

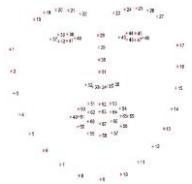
    → GPU server (*Facebook AI Research Partnership*)

1. *Clive Humby, founder of the Clubcard, the world’s first supermarket loyalty scheme*

Collected Datasets

- **Pandora**
  - Goal: *Head and Shoulder Pose Estimation* task
  - **250k** images (**Full HD RGB** and **depth**) of the upper body
  - **22 subjects**
  - Annotations: **head** and **shoulder** angles (*yaw*, *pitch* and *roll*)
    - **Head**: ±70° roll, ±100° pitch and ±125° yaw
    - **Shoulder**: ±70° roll, ±60° pitch and ±60° yaw
  - Challenging **camouflage** (glasses, scarves, caps…)

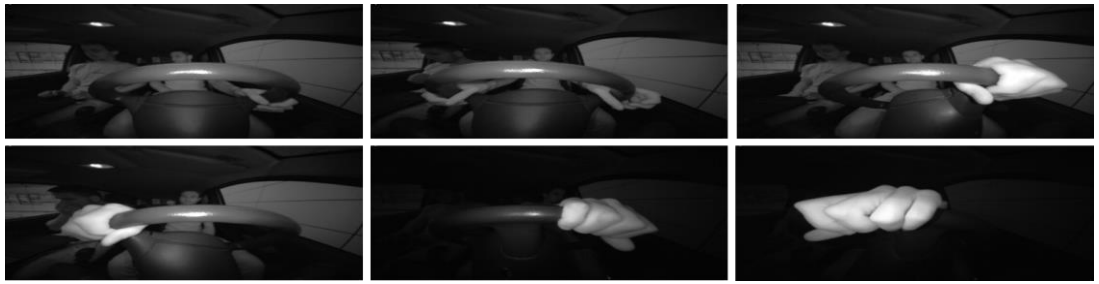
- **MotorMark**
  - Goal: *Facial Landmark Localization* task
  - **30k** images (RGB and depth) of the upper body
  - **35 subjects** with head garments
  - Real automotive context
  - Facial Landmark annotations following the *ISO MPEG-4* standard



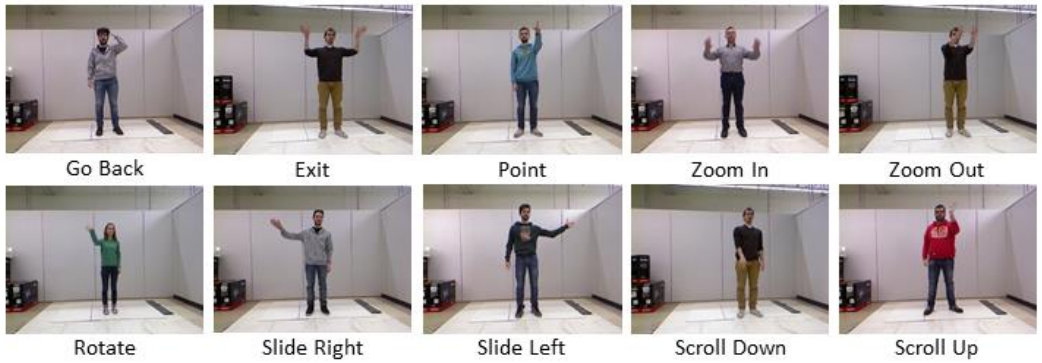
1. *Pandora dataset: http://imagelab.ing.unimore.it/pandora/*
2. *MotorMark dataset: http://imagelab.ing.unimore.it/motormark-dataset*

Collected Datasets

- **Turms**
  - Goal: *Driver's hands detection and tracking*
  - **14k infrared** (*IR*) frames acquired
  - *Leap Motion device* (**stereo camera**)
  - **Annotations**: bounding boxes of right and left hands
  - Original position: back to the steering wheel



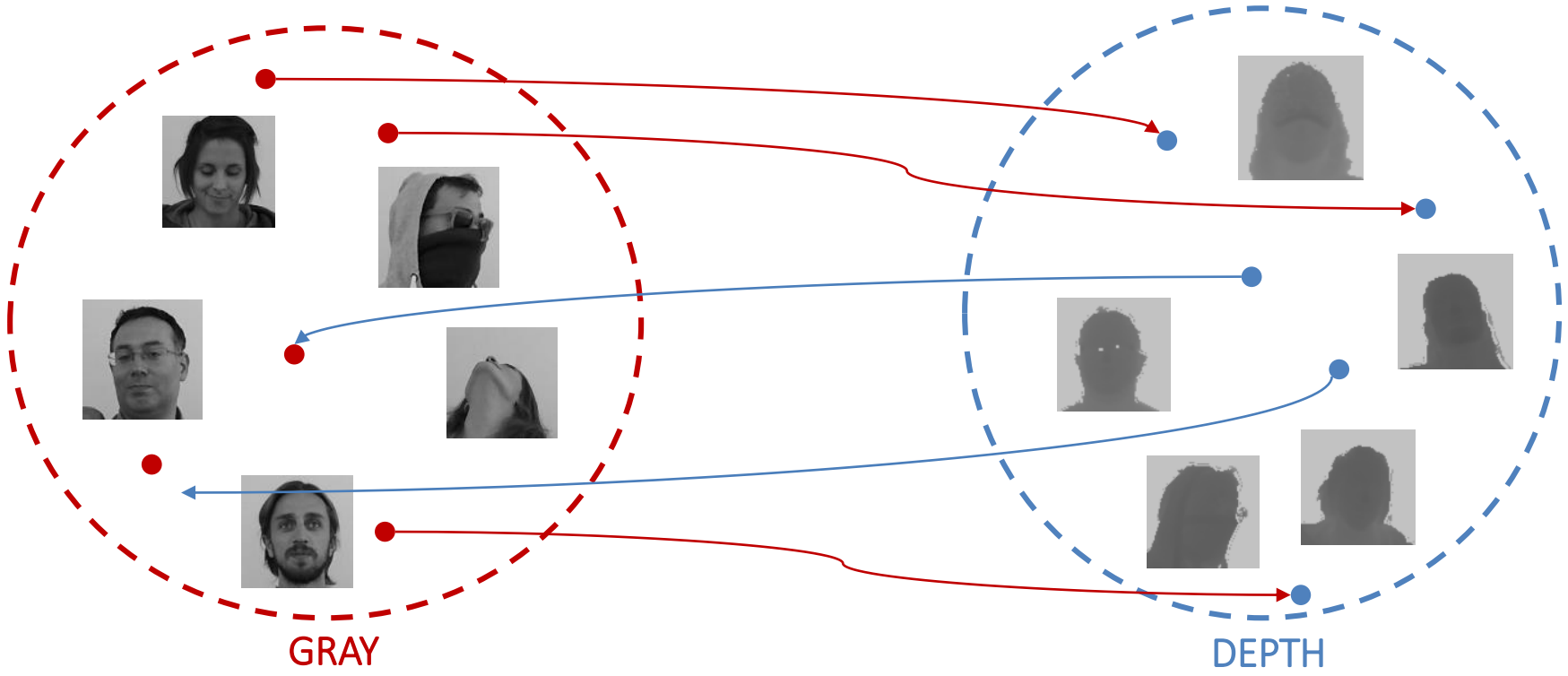
- **Kinteract**
  - Goal: *Human-Computer Interaction through Body Gestures*
  - **168 sequences** of gestures
  - RGB frames and body joints on depth maps
  - **10 subjects**
  - Classes: *Go Back, Exit, Point, Zoom In/Out, Rotate, Slide Right/Left, Scroll Down/Up*



1. *Turms: http://imagelab.ing.unimore.it/turms*
2. *Kinteract: http://imagelab.ing.unimore.it/hci*

Face Translation: *Face-from-Depth*

Is it possible to *generate* gray-level face images from the corresponding depth ones?

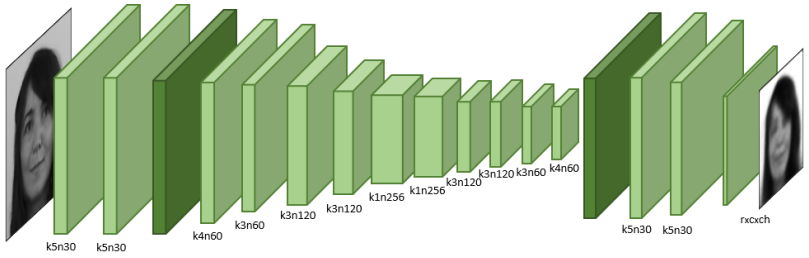


1. *G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara "Face-from-depth for head pose estimation on depth images", TPAMI 2018*
2. *G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara "Poseidon: Face-from-depth for driver pose estimation", CVPR 2017*
3. *M. Fabbri, G. Borghi, F. Lanzi, R. Vezzani, S. Calderara, and R. Cucchiara. "Domain translation with conditional gans: from depth to rgb face-to-face", ICPR 2018*

Face Translation: *Face-from-Depth*

- The main idea is to **add knowledge** (the *generated* faces) at training and testing time, **to improve the performance**
- We propose (two versions) of a **new** neural network called *Face-from-Depth*:

FfD [v1]: CNN (CVPR 17)



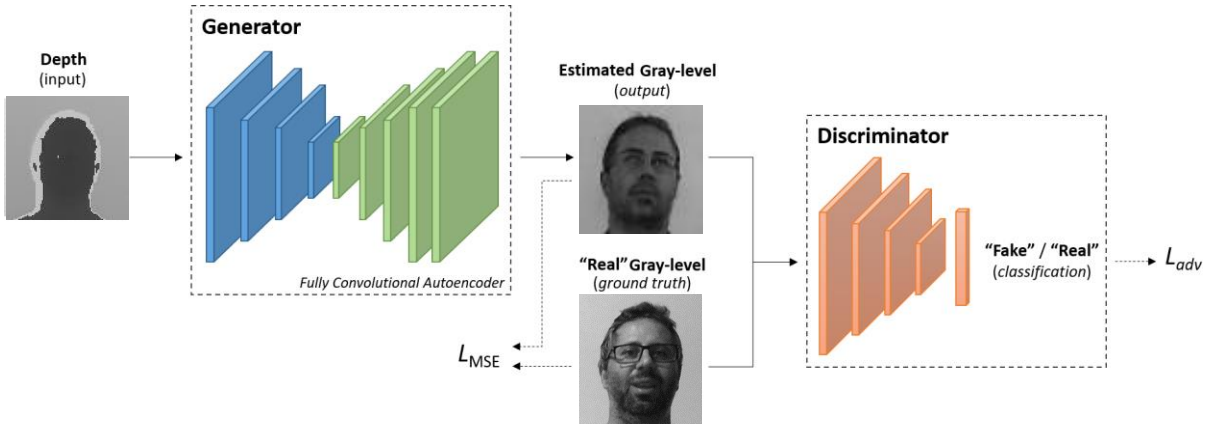
- Elements from **autoencoder** and **CNN**
- **Weighted Loss**:

$$L = \frac{1}{R\,C}\sum_{i}^{R}\sum_{j}^{C}(|y_{ij} - y'_{ij}|_2^2 \cdot w_{ij}^{\mathcal{N}})$$

$$\mathcal{N}:\ \mu = \left[\frac{R}{2}, \frac{C}{2}\right]^T \quad \sum = \mathbb{I} \cdot \left[\left(\frac{R}{\alpha}\right)^2, \left(\frac{C}{\beta}\right)^2\right]^T \qquad \alpha = 3.5, \beta = 2.5$$

- *Bi-variate Gaussian* prior mask to highlight the central area

FfD [v2]: *conditional GAN* (PAMI 18)



- **Min-max game** (Generator - Discriminator):

$$\min_{\theta_d}\max_{\theta_g}\mathbb{E}_{x\sim p_{gray}(x)}[\log(G(x))] + \mathbb{E}_{y\sim p_{dpt}(y)}[\log(1 - G(D(y)))]$$

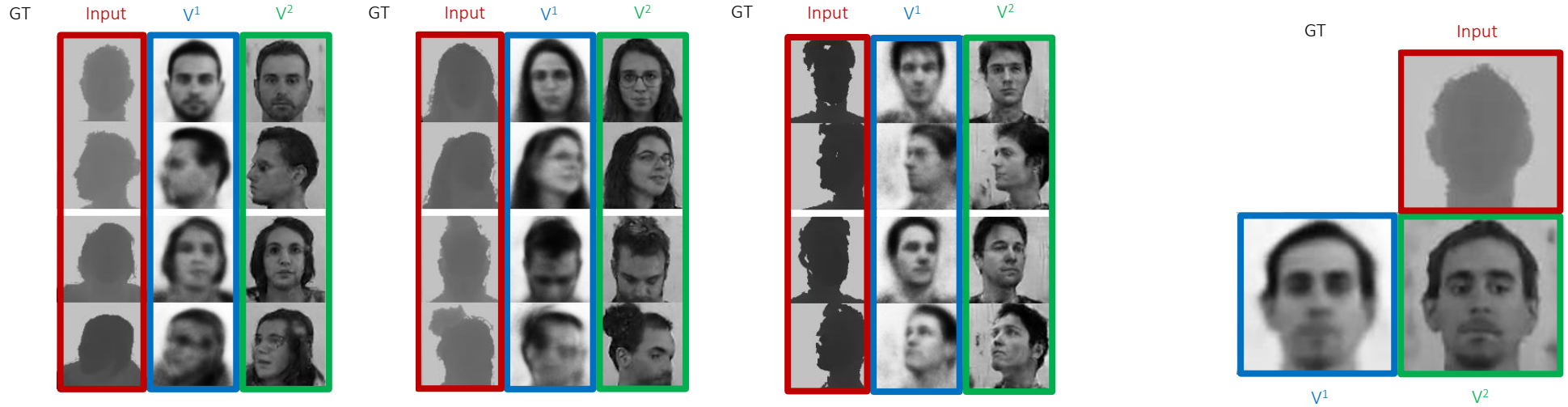
- **2 loss functions**:

$$L_{MSE}(s^g, s^d) = \frac{1}{N}\sum_{i=1}^{N}|G(s_i^g) - s_i^d|_2^2$$

$$L_{adv}(y, t) = -\frac{1}{N}\sum_{i=1}^{N}[t_i \log y_i + (1 - t_i)\log(1 - y_i)]$$

Face Translation: *Face-from-Depth*

- **Visual** and **numerical** comparisons between FfD [v1] and FfD [v2]:



Pandora dataset

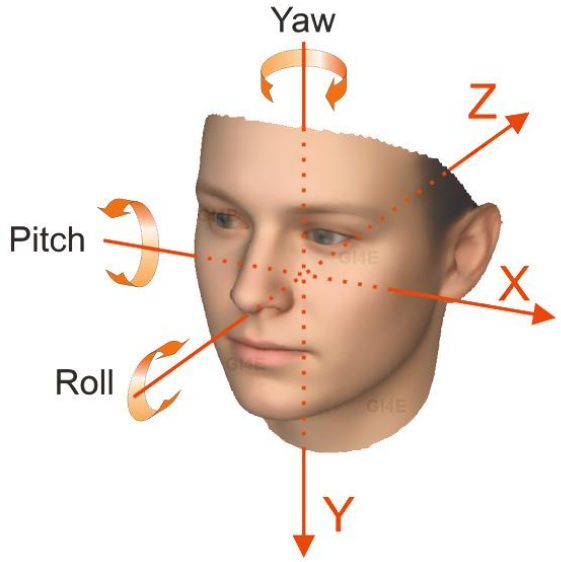
Biwi dataset

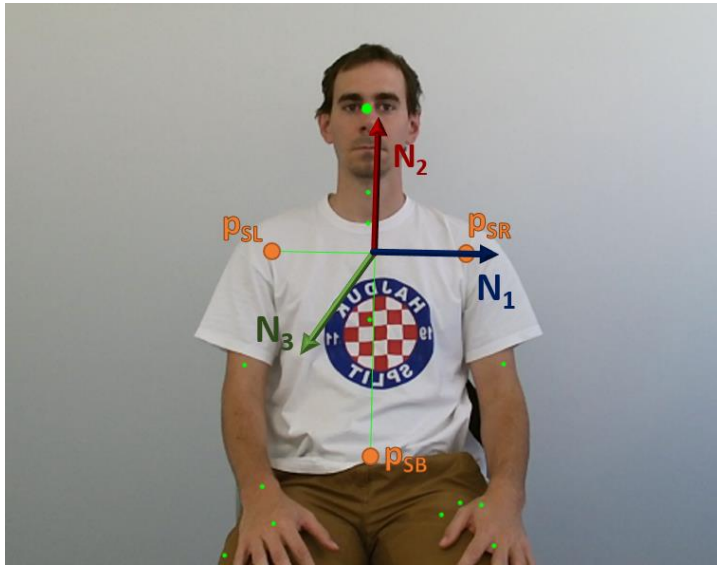
Dataset	Method	Norm ↓		Difference ↓		RMSE ↓			Threshold ↑		
		L_1	L_2	Abs	Squared	linear	log	scale-inv	1.25	2.5	3.75
Biwi	FfD[1]	33.35	2586	0.454	24.07	40.55	**0.489**	**0.445**	0.507	**0.806**	**0.878**
	FfD	**24.44**	**2230**	**0.388**	**19.81**	**35.50**	0.653	0.610	**0.615**	0.764	0.840
Pandora	FfD[1]	41.36	3226	0.705	46.00	50.77	0.603	**0.485**	0.263	0.725	0.819
	pix2pix[2]	19.37	1909	**0.468**	24.07	30.80	0.568	0.539	0.583	0.722	0.813
	AVSS[3]	23.93	2226	0.629	34.49	35.46	0.658	0.579	0.541	0.675	0.764
	FfD + U-Net	23.75	2123	0.653	34.96	33.89	0.639	0.553	0.555	0.689	0.775
	FfD	**18.21**	**1808**	0.469	**22.90**	**28.90**	**0.556**	0.501	**0.605**	**0.743**	**0.828**

1. *G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation", CVPR 2017*
2. *P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks", CVPR 2017*
3. *M. Fabbri, S. Calderara, and R. Cucchiara, "Generative adversarial models for people attribute recognition in surveillance", AVSS 2017*

Head and Shoulder Pose Estimation

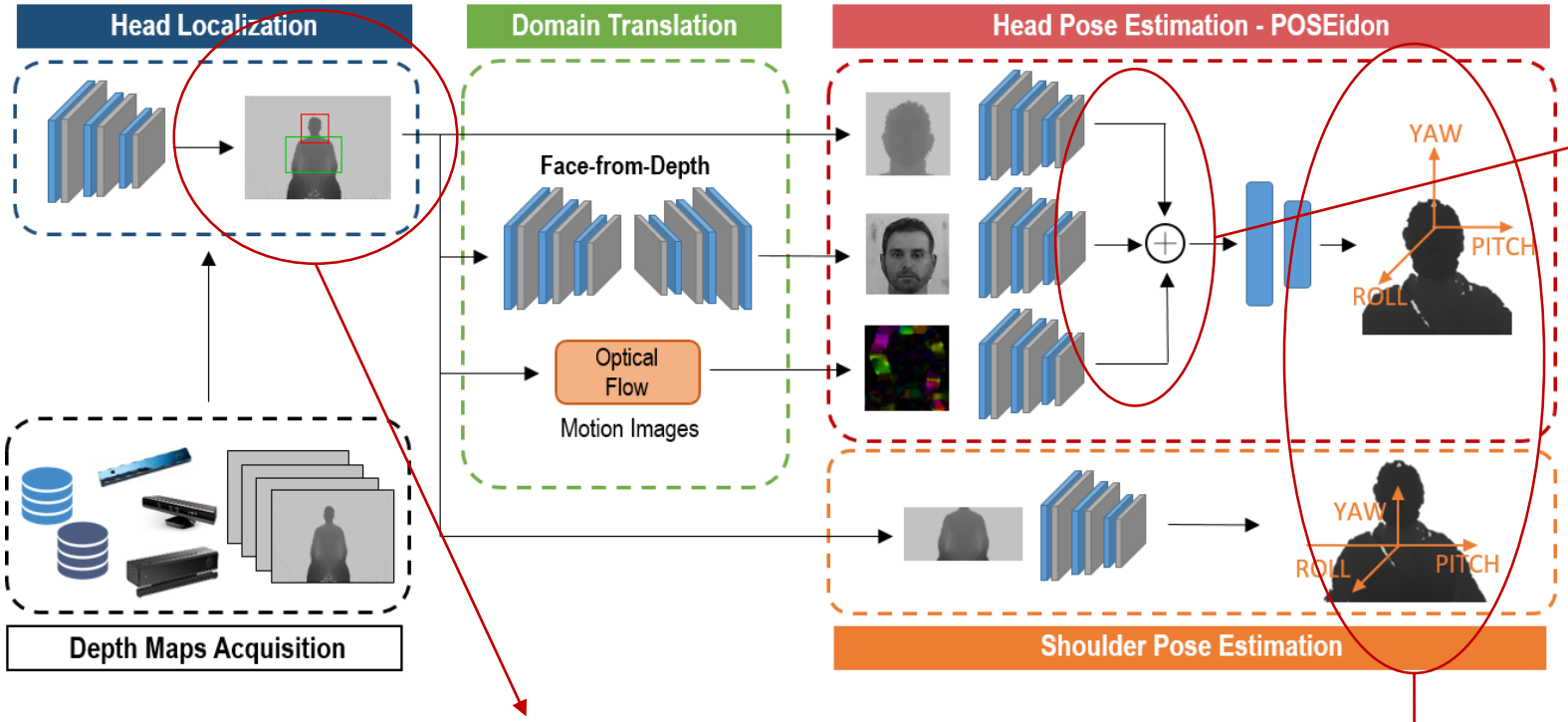
It is the ability to infer the orientation of the person's head (shoulders) relative to the view of a camera





1. *G. Borghi, R. Gasparini, R. Vezzani, and R. Cucchiara, "Embedded recurrent network for head pose estimation in car", IV 2017*
2. *M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara. "From depth data to head pose estimation: a siamese approach", VISAPP 2017*
3. *M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara, "Deep head pose estimation from depth data for in-car automotive applications", ICPRW 2016*
4. *G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara "Poseidon: Face-from-depth for driver pose estimation", CVPR 2017*

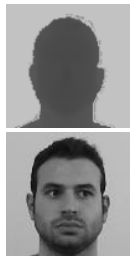
POSEidon: Body Pose Estimation [CVPR 2017, VISAPP 2017, IV 2017 , ICPRW 2016]



Head crop formula

$$w, h = \frac{f_{x,y} \cdot R}{D}$$

- $f_{x,y}$ are the horizontal and vertical focal lengths
- R is the average value representing the width of a face (200mm)
- D is the distance between the acquisition device and the head in the scene.



Final Outputs

- 3D angles *yaw*, *pitch* and *roll* for head and shoulders

Training Procedure

- Double-step procedure
  - 1st: each individual network is trained
  - 2nd: the last fc layer is removed, networks are then merged through a *conv* and *concat* operations:

$$y^{cat} = [x^a | x^b], \qquad d^y = d_a^x + d_b^x$$

$$y^{cnv} = y^{cat} * k + \beta, d^y = \frac{(d_a^x + d_b^x)}{2}$$

x^a , x^b: feature maps
d_a^x , d_b^x: feature channel

- Loss function:

$$L = \sum_{i=1}^{3} |w_i \cdot (y_i - f(x_i))|_2^2$$

$$\boldsymbol{w} = [0{,}2, 0{,}35, 0{,}45]$$

POSEidon: Body Pose Estimation [CVPR 2017, VISAPP 2017, IV 2017 , ICPRW 2016]

Validation Procedure	Year	Data Depth	Data RGB	Head Pitch	Head Roll	Head Yaw	Avg
ALL SEQUENCES USED AS TEST SET							
Padeleris [38]	2012	√		6.6	6.7	11.1	8.1
Rekik [55]	2013	√	√	4.3	5.2	5.1	4.9
Martin [72]	2014	√		2.5	2.6	3.6	2.9
Papazov [37]	2015	√		2.5 ± 7.4	3.8 ± 16.0	3.0 ± 9.6	4.0 ± 11.0
Meyer [8]	2015	√		2.4	2.1	2.1	2.2
Li [54]	2016	√	√	1.7	3.2	2.2	2.4
Sheng [41]	2017	√		2.0	1.9	2.3	2.1
LEAVE ONE OUT (LOO)							
Drouard [12]	2015		√	5.9 ± 4.8	4.7 ± 4.6	4.9 ± 4.1	5.2 ± 4.5
Drouard [19]	2017		√	10.0 ± 8.7	8.4 ± 8.0	8.6 ± 7.2	9.0 ± 7.9
POSEidon$^+$	2017	√		**2.4 ± 1.3**	**2.6 ± 1.5**	**2.9 ± 1.5**	**2.6 ± 1.4**
K4-FOLD SUBJECT CROSS VALIDATION							
Fanelli [35]	2011	√		3.5 ± 5.8	5.4 ± 6.0	3.8 ± 6.5	- ± -
POSEidon$^+$	2017	√		**2.8 ± 1.7**	**2.9 ± 2.1**	**3.6 ± 2.5**	**3.1 ± 2.1**
K5-FOLD SUBJECT CROSS VALIDATION							
Fanelli [34]	2011	√		8.5 ± 9.9	7.9 ± 8.3	8.9 ± 13.0	8.43 ± 10.4
POSEidon$^+$	2017	√		**2.8 ± 1.8**	**2.8 ± 2.2**	**3.6 ± 2.2**	**3.0 ± 2.1**
K8-FOLD SUBJECT CROSS VALIDATION							
Lathuiliere [25]	2017		√	4.7	3.1	**3.1**	3.6
POSEidon$^+$	2017	√		**2.8 ± 1.9**	**2.8 ± 1.8**	3.3 ± 2.0	**3.0 ± 1.9**
FIXED TRAIN AND TEST SPLITS							
Yang [44]	2012	√	√	9.1 ± 7.4	7.4 ± 4.9	8.9 ± 8.3	8.5 ± 6.9
Baltrusaitis [50]	2012	√	√	5.1	11.3	6.3	7.6
Kaymak [47]	2013	√	√	7.4	6.6	5.0	6.3
Wang [73]	2013	√	√	8.5 ± 14.3	7.4 ± 10.8	8.8 ± 14.3	8.2 ± 12.0
Ahn [11]	2014		√	3.4 ± 2.9	2.6 ± 2.5	2.8 ± 2.4	2.9 ± 2.6
Saeed [45]	2015	√	√	5.0 ± 5.8	4.3 ± 4.6	3.9 ± 4.2	4.4 ± 4.9
Liu [27]	2016		√	6.0 ± 5.8	5.7 ± 7.3	6.1 ± 5.2	5.9 ± 6.1
POSEidon [10]	2017	√		1.6 ± 1.7	1.8 ± 1.8	1.7 ± 1.5	1.7 ± 1.7
POSEidon$^+$	2017	√		**1.6 ± 1.3**	**1.7 ± 1.7**	**1.7 ± 1.3**	**1.6 ± 1.4**

Accuracy on *Biwi* dataset.
POSEidon overcomes all the competitors using different evaluation procedures.

#	Input Depth	Input FfD	Input MI	Input Gray	Crop	Fusion	Head Pitch	Head Roll	Head Yaw	Accuracy
1	√					-	8.1 ± 7.1	6.2 ± 6.3	11.7 ± 12.2	0.553
2	√				√	-	6.5 ± 6.6	5.4 ± 5.1	10.4 ± 11.8	0.646
3		√			√	-	6.8 ± 6.1	5.8 ± 5.0	10.1 ± 12.6	0.658
4			√		√	-	7.7 ± 7.5	5.3 ± 5.7	10.0 ± 12.5	0.609
5				√	√	-	7.1 ± 6.6	5.6 ± 5.8	9.0 ± 10.9	0.639
6	√	√			√	concat	5.6 ± 5.0	4.9 ± 5.0	9.7 ± 12.1	0.698
7	√		√		√	concat	6.0 ± 6.1	4.5 ± 4.8	9.2 ±11.5	0.690
8	√	√	√		√	conv+concat	**5.6 ± 5.2**	**4.8 ± 5.0**	**8.2 ± 9.8**	**0.736**

Accuracy on *Pandora* dataset (head)

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**

Accuracy on *Pandora* dataset (shoulder)

1. *Too many references, please see the original paper*
   G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara "Face-from-depth for head pose estimation on depth images" PAMI 2018

Face Verification

Comparison of two faces in order to determine whether they belong to the same person or not (*one-to-one* comparison)

It is different from *Face Identification* that is a comparison of a unknown subject's face with a set of known faces


Face Verification from Depth using Privileged Information

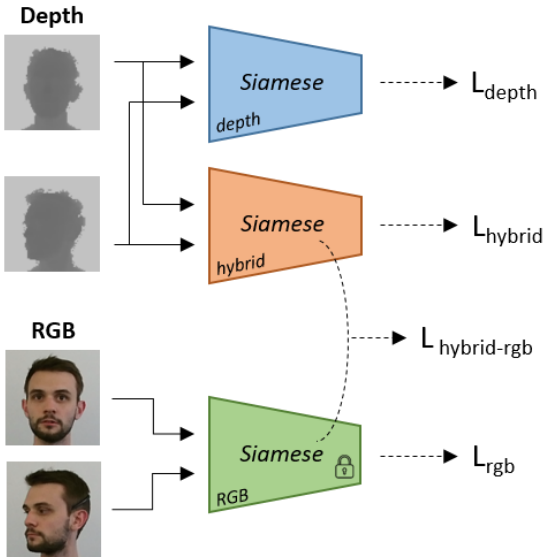

Driver Face Verification with Depth Maps

1. *G. Borghi, S. Pini, F. Grazioli, R. Vezzani, R. Cucchiara, ''Face Verification from Depth using Privileged Information'', BMVC 2018*
2. *G. Borghi, S. Pini, F. Grazioli, R. Vezzani, R. Cucchiara , ''Driver Face Verification with Depth Maps'', submitted to Sensors Journal*

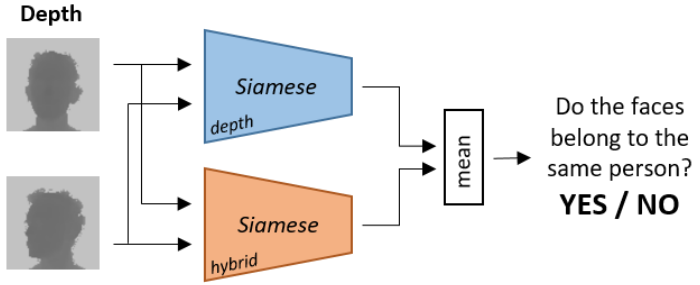
JanusNet: Face Verification [BMVC18]

- Privileged Information: additional (privileged) knowledge available only during the training but it improves the performance of the system at testing time[1, 2]
- We propose the JanusNet system: *Siamese* networks + 3 models (train) + 2 models (test)

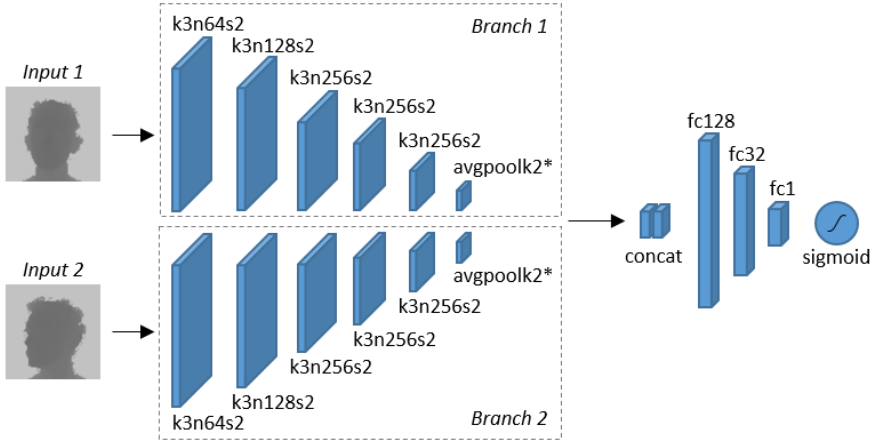
Train



Test



Siamese Network



Privileged Information Loss:

$$L_{hybrid-rgb_{1,2}} = \frac{1}{N}\sum_{n}^{N}(y_n^{hybrid} - y_n^{rgb})^2$$

Final Loss:

$$L = \alpha\left(L_{hybrid-rgb_1} + L_{hybrid-rgb_2}\right) + \beta\left(L_{depth} + L_{hybrid} + L_{rgb}\right)$$

1. Vapnik et al., "A new learning paradigm: Learning using privileged information", Neural Networks, 2009.
2. Hoffman et al., "Learning with Side Information through Modality Hallucination", IEEE CVPR, 2016.

JanusNet: Face Verification [BMVC18]

- Accuracy on Face Verification
  - JanusNet architecture **outperforms the single Siamese models**.
  - The face verification accuracy of the proposed model is **comparable to a well-known deep architecture** pre-trained on RGB images.

Model	Data type	Accuracy
FaceNet	RGB	0.8232
Hybrid network	Depth	0.7553
RGB network	RGB	0.7631
Depth network	Depth	0.7950
JanusNet	**Depth**	**0.8142**

- Accuracy on Face Identification
  - To deal with the *one-to-many* comparison, JanusNet is used to obtain **a similarity score between every possible pair of images contained in the dataset**.
  - We combine the results with the following functions:

$$\mathrm{y} = \underset{i}{\operatorname{argmax}}\, J(s, s'),\ \forall s' \in S_i \qquad \mathrm{y} = \underset{i}{\operatorname{argmax}}\ \underset{s' \in S_i}{\operatorname{avg}}\, J(s, s') \qquad \mathrm{y} = \underset{i}{\operatorname{argmax}}\, \#\{S_i \mid J(s, s') > t\},\ \forall s' \in S_i$$

	Pegasos SVM				JanusNet		
	LBP	SIFT	DLQP	Bag-D3P	*max*	*avg*	*voting*
Accuracy	0.5917	0.7194	0.7347	0.9430	0.9756	**0.9877**	0.9804
Improvement	-	+12.7	+14.3	+35.1	+38.4	**+39.6**	+38.9

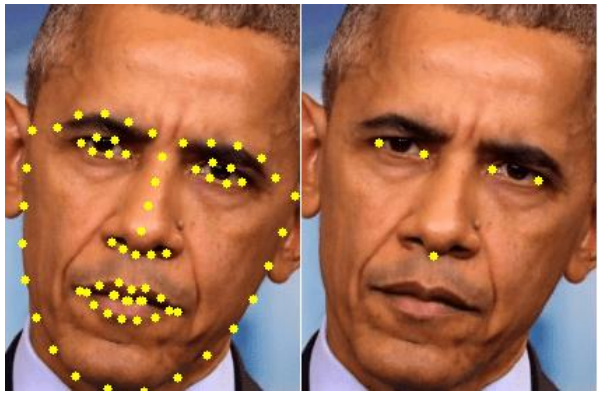
Competitor results are taken from:

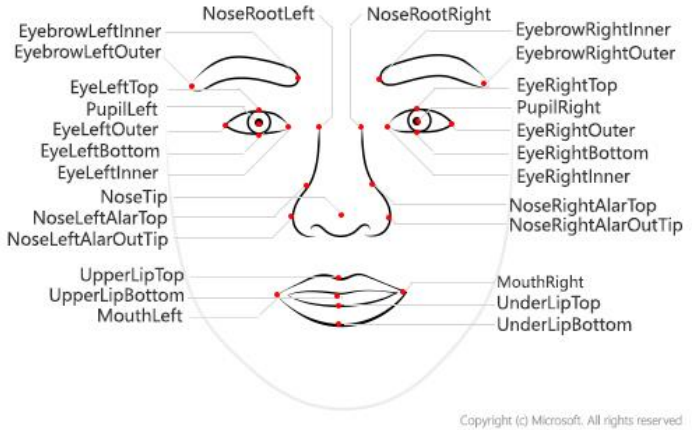
T. Mantecon et al. "Depth-based face recognition using local quantized patterns adapted for range data", ICIP, 2014.

T. Mantecon et al. " Visual face recognition using bag of dense derivative depth patterns", IEEE SPL, 2016.

Facial Landmark Localization

Detection of the position of salient points of the human face.
In our case: eyes' pupils, mouth corners and the nose tip.



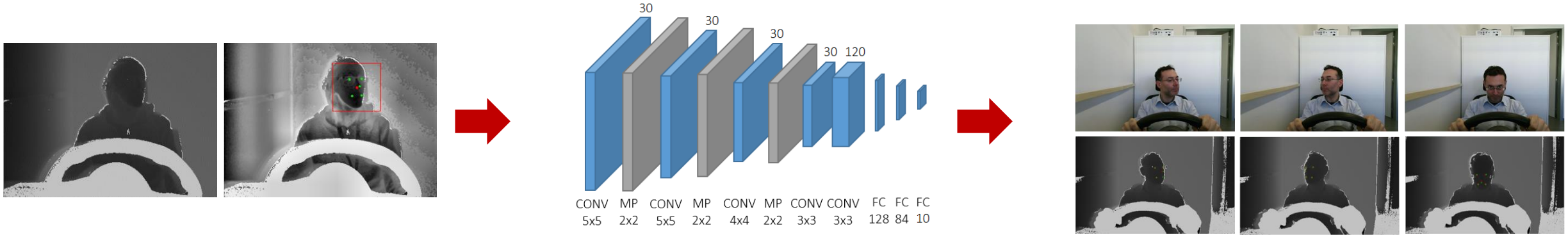



Fast and Accurate Facial Landmark Localization in Depth Images

1. *E. Frigieri, G. Borghi, R. Vezzani, and R. Cucchiara, "Fast and accurate facial landmark localization in depth images for in-car applications", ICIAP 2017*

Facial Landmark Localization [ICIAP 2017]

- Proposed pipeline:



Input pre-processing

- *Contrast Limited Adaptive Histogram Equalization* algorithm applied
- Values scaled:
  mean = 0, *variance* = 1
- A fixed window containing the head is cropped and all the cropped images are resized to 64x64 pixels

Deep Network

Shallow model (real time performance)

- 5 *convolutional* layers
- 3 *max-pooling* layers
- 3 *fully connected* layers
- Activation function: *tanh*
- L_2 loss:

$$L_2 = \sum_{i=1}^{3} ||(y_i - f(x_i))||_2^2$$

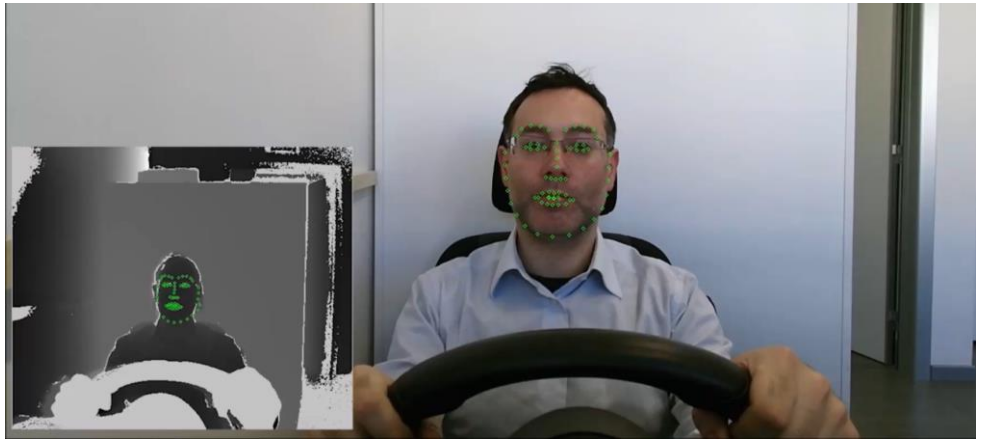
Output

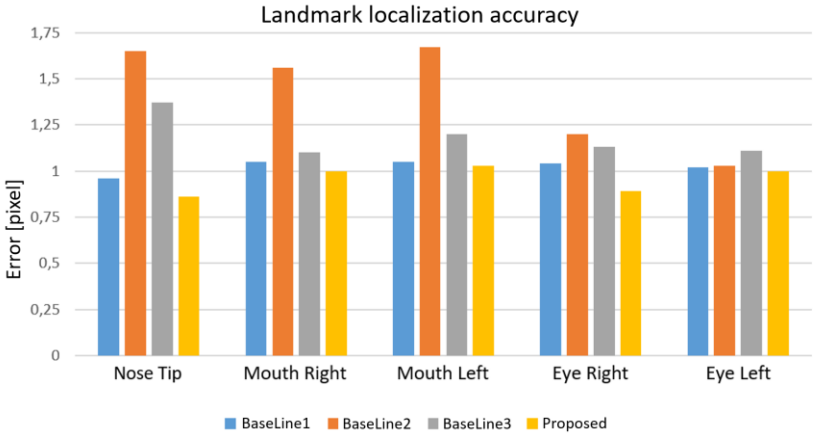
- **10 coordinates** (*x,y*) of 5 facial landmarks
- Ground truth coordinates are normalized in the range [-1, 1] accordingly to the specific activation function of the output network layer

Facial Landmark Localization [ICIAP 2017]

Method	Nose Tip	Mouth Right	Mouth Left	Eye Right	Eye Left	Avg Err
Zhao *et al.*	4.4±2.2	5.4±3.2	5.4±3.2	4.2±2.1	4.2±2.2	4.7±2.6
Our	3.3±4.5	3.5±3.7	3.4±3.9	3.5±4.1	3.4±4.0	3.4±4.0

Results on *Eurecom Kinect Face dataset*, expressed as the mean error and the standard deviation in pixels w.r.t the ground truth, normalized by the interpupillary distance.





Results on *MotorMark* dataset expressed as the mean error and the std in pixels w.r.t the ground truth, normalized by the interpupillary distance.

Different tests have been carried out:

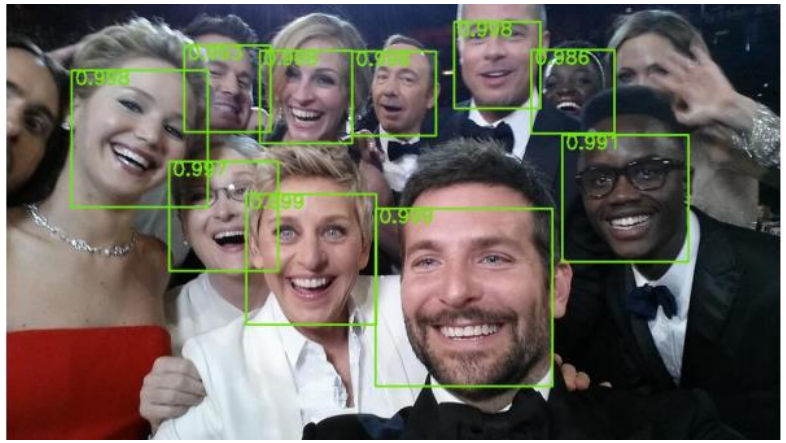
- Smaller window cropped from head center
- Bigger input images (128x128)
- Background suppression applied
- The proposed method

1. Zhao, Xi, et al. "Automatic 2.5-D facial landmarking and emotion annotation for social interaction assistance." IEEE transactions on cybernetics (2016)

Head/Face Detection

It is the ability to detect and localize one or more heads/faces in an image.

This is a traditional problem of the computer vision field,
but only few works tackle this task on different types of images, like depth maps.



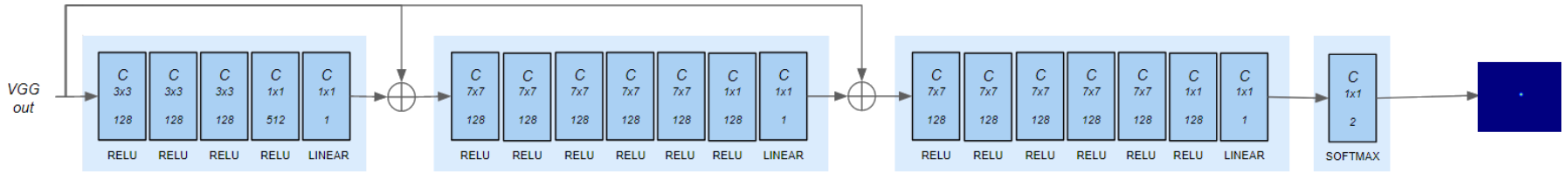

Fully Convolutional Network for Head Detection with Depth Images


Head Detection with Depth Images in the Wild

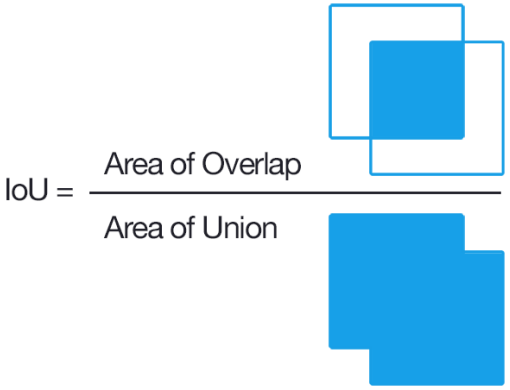
1. *D. Ballotta, G. Borghi, R. Vezzani, and R. Cucchiara, ''Head detection with depth images in the wild'', VISAPP 2017*
2. *D. Ballotta, G. Borghi, R. Vezzani, R. Cucchiara, "Fully Convolutional Network for Head Detection with Depth Images'', ICPR 2018*

Head Detection

- Proposed method:



  - **Input**: depth maps (512x424 16-bit images, from *Microsoft Kinect One*)
  - **Output**: 64x53 probability map (as *bi-variate Gaussian* function)
  - **Network Architecture**: Fully Convolutional Network (inspired by CPM [1])
    - We **modify** the original architecture and **reduce** the number of parameters to deal with:
      - Real time performance
      - Lack of training data
  - **Network Details**: (*ReLU* + linear activation for each block) + *softmax*
  - **Loss function**: categorical cross-entropy



The final accuracy is evaluated through the *Intersection over Union* metric:

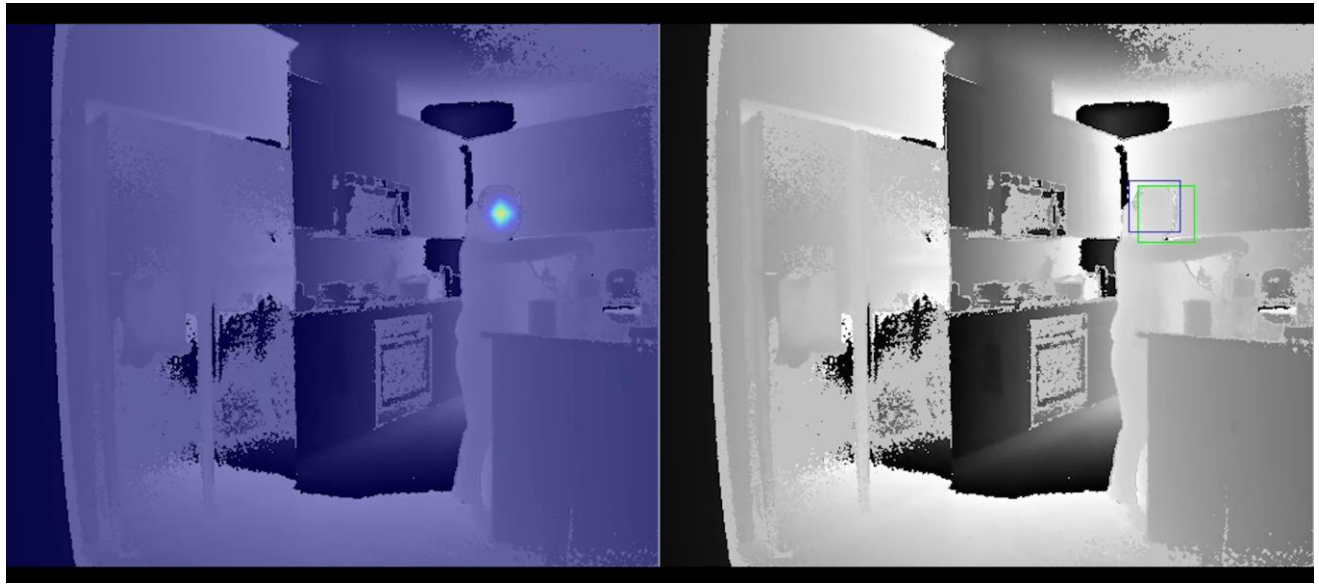
$$IoU(A,B) > 0.5 \qquad IoU(A,B) = \frac{Overlap\ Area}{Union\ Area} = \frac{|A \cap B|}{|A \cup B| - |A \cap B|}$$

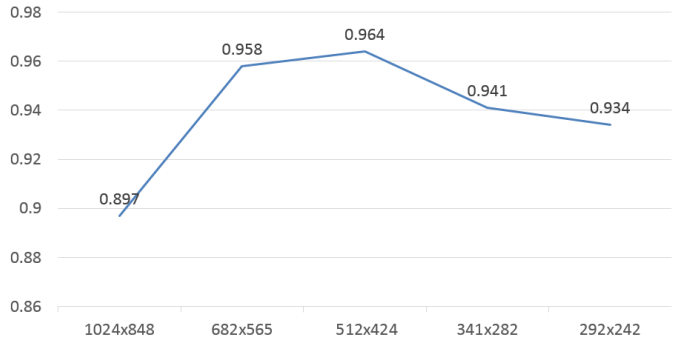
1. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines" (CVPR 2016)

Head Detection

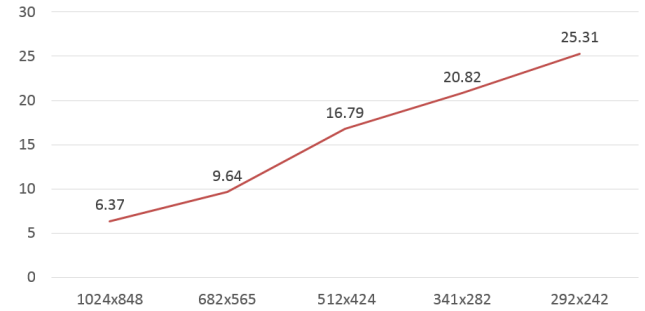
Methods	Year	Method	TP	FP
Nghiem *et al.* [3]	2012	SVM	0.519	0.076
Chen *et al.* [4]	2016	LDA	0.709	0.108
Ballotta *et al.* [5]	2017	CNN	0.883	0.077
Our	**2018**	**CNN**	**0.964**	**0.036**

Result on *Watch-n-Patch* dataset





Rate of true head detection over the change of input shape



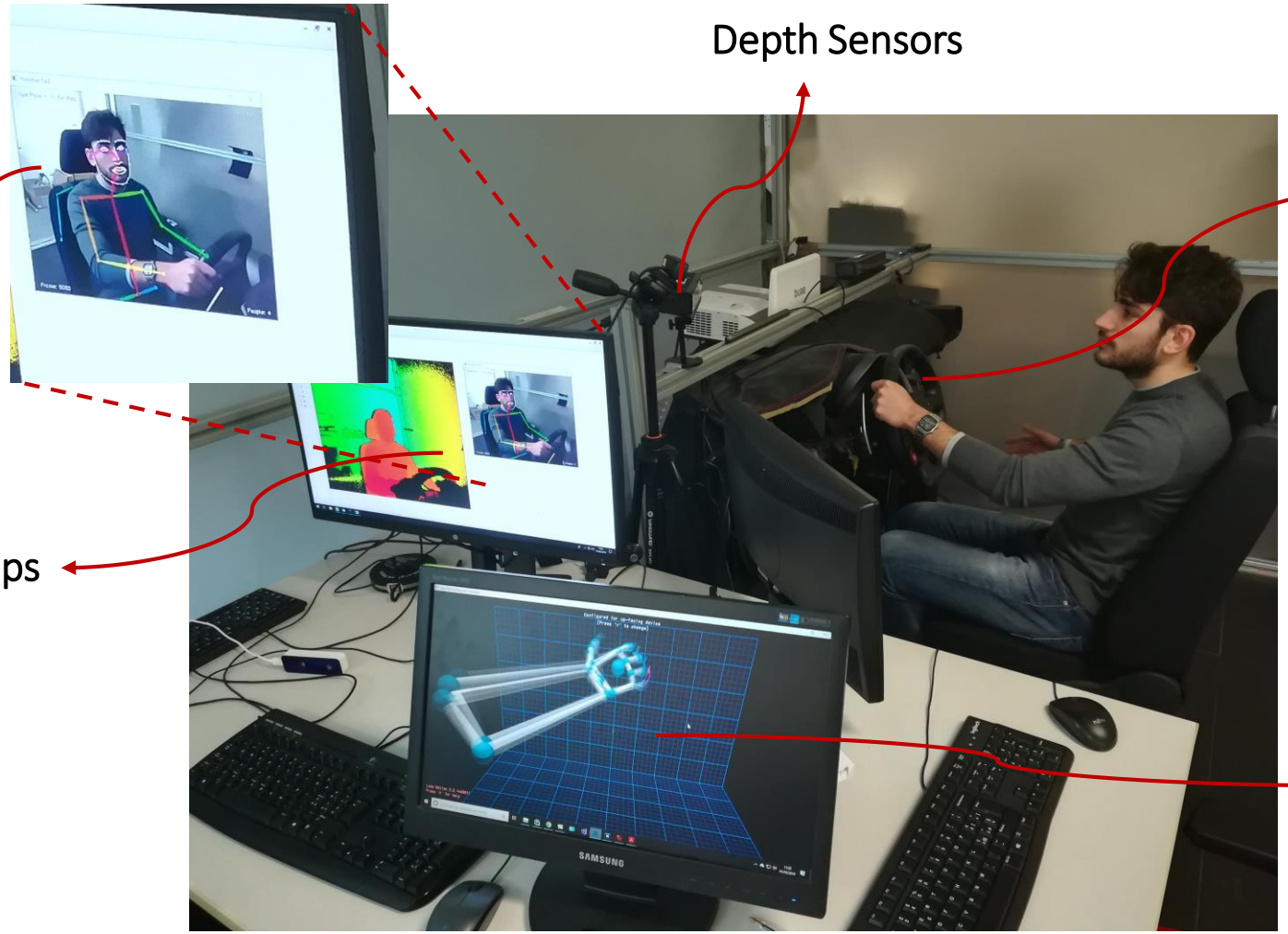
Speed performance over the change of input shape (*fps*)

[3] A. T. Nghiem, E. Auvinet, and J. Meunier, "Head detection using kinect camera and its application to fall detection" (ISSPA 2012)

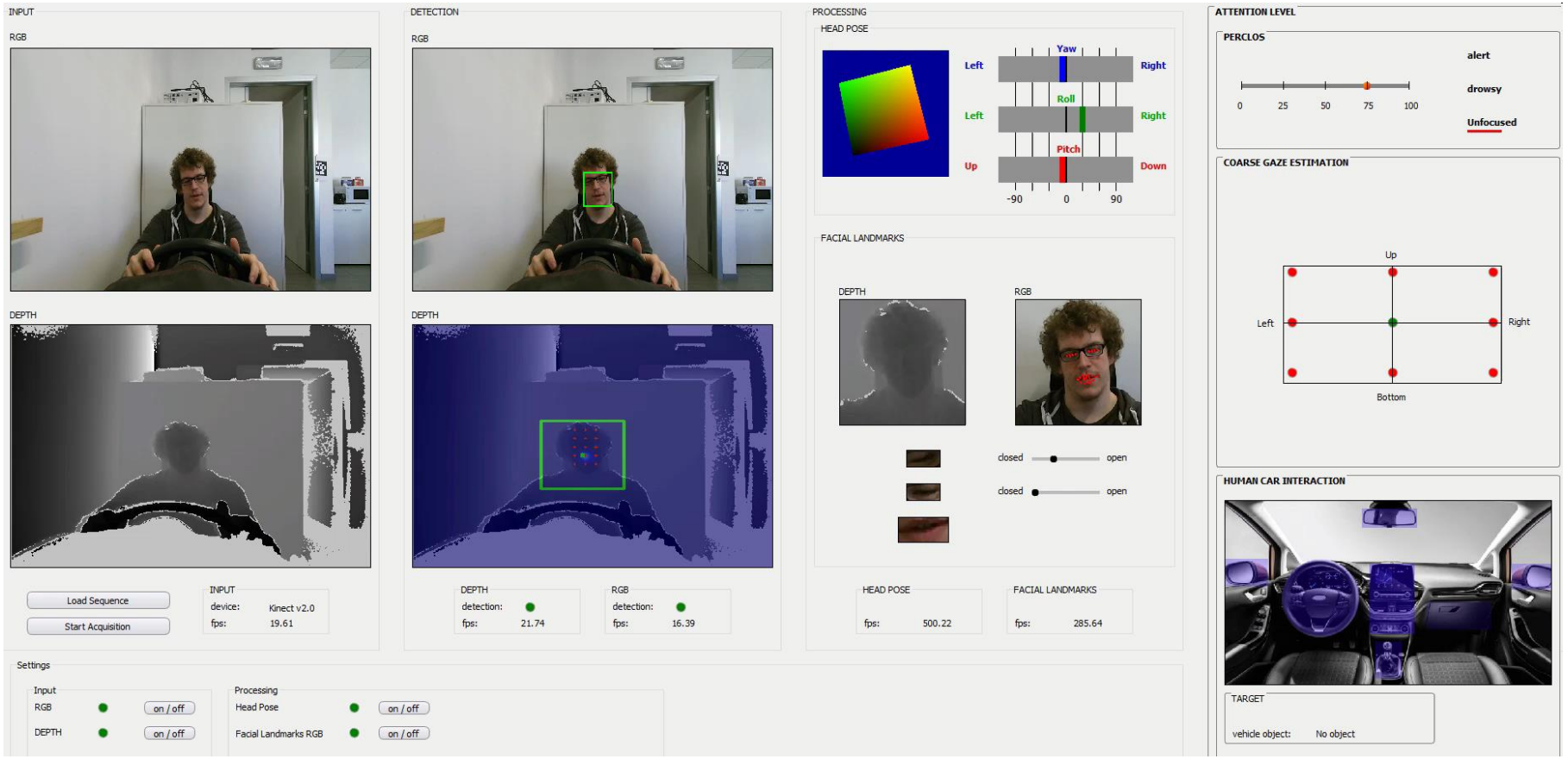
[4] S. Chen, F. Bremond, H. Nguyen, and H. Thomas, "Exploring depth information for head detection with depth images" (AVSS 2016)

[5] D. Ballotta, G. Borghi, R. Vezzani, R. Cucchiara, "Head Detection with Depth Images in the Wild" (VISAPP 2017)

Data Acquisition in AImagelab



Mercury



Perclos
Driver drowsiness level

Coarse Gaze Est.
Where is the driver looking at?

Driver-Car Interaction
What object is the driver looking at?

Microsoft Kinect
(RGB + Depth)

Head Detection
(on RGB and Depth)

Head Pose Estimation
Facial Landmarks

Driver's Attention
Eye State