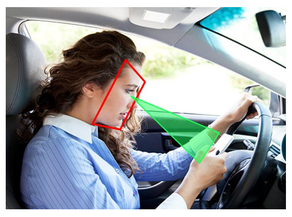
EMBEDDED RECURRENT NETWORK FOR HEAD POSE ESTIMATION IN CAR

Guido Borghi, Riccardo Gasparini, Roberto Vezzani, Rita Cucchiara
University of Modena and Reggio Emilia, Italy
name.surname@unimore.it

The 28th IEEE IV2017 June 11-14, Redondo Beach, CA, USA

Abstract

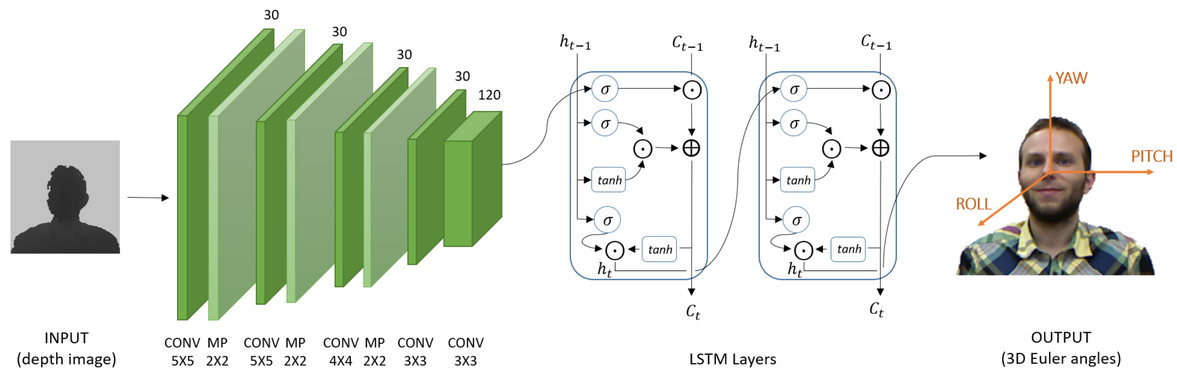


An accurate and fast **driver's head pose estimation** is a rich source of information, in particular in the automotive context.
Head pose is a key element for driver's behavior investigation, pose analysis, attention monitoring and also a useful component to improve the efficacy of Human-Car Interaction systems.

In this paper, a **Recurrent Neural Network** is exploited to tackle the problem of driver head pose estimation, directly and only working on **depth images to be more reliable in presence of varying or insufficient illumination**.

Experimental results, obtained from two public dataset, namely *Biwi Kinect Head Pose* [1] and *ICT-3DHP Database* [2], prove the efficacy of the proposed method that overcomes state-of-art works.
Besides, the entire system is implemented and tested on two **embedded boards with real time performance**.

Proposed System



Input: *depth* frame; user's head is cropped with a *dynamic* window (*w, h*) to include smaller part of the background:

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

$f_{x,y}$: focal length , R: width of a generic face, Z: distance between the subject's face and the device.

Output: 3D continuous pose angle (*yaw*, *pitch* and *roll*)

- RNN is exploited in a **regression** manner
- Ground truth angles normalized between [-1, +1]

How:

- Recurrent Neural Network (*Long-Short Term Memory*, LSTM)
- Temporal sequences of 60 frames (64x64 pixels)

Training details:

- Optimizer: *Adadelta*
- Activation function: *Hyperbolic Tangent*, to map values $[-\infty, +\infty] \rightarrow [-1, +1]$
- Loss function: $L_2 = \sum_1^n ||y_i - f(x_i)||_2$

Introduction

Driver head pose estimation can be a fundamental element to monitor:

- **Distraction**
  - Visual: driver's gaze is not on the road
  - Cognitive: driver is not focused
  - Manual: hands are away from steering wheel
- **Fatigue**
  - Local: skeletal of ocular fatigue
  - General: consequence of a manual labor
  - Central Nervous: drowsiness
  - Mental: low concentration

And to develop new **Human-Car Interaction** systems:

- User-friendly (Natural Body Language)
- Fast
- Safe (no hand contact is required)

Automotive context **requirements**:

- Light invariance (night/day, weather conditions)
- No invasivity
- Real time performance
- Direct estimation from input images

Our **solutions**:

- Infrared sensors (depth device)
- Computer Vision algorithms
- Shallow deep architecture
- No facial features (landmarks, nose tip…)

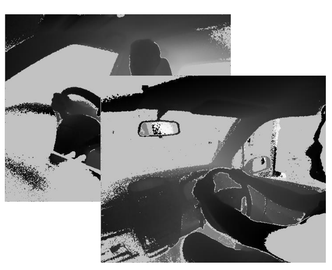


Figure 1. Example of depth frames acquired in automotive context

Video Demo

Head Pose Results

The first public dataset (*Biwi Kinect Head Pose Dataset*), is used for the training phase. An additional dataset (*ICT-3D database*) is included in the testing phase, **conducing a cross-dataset evaluation**.

The evaluation metric is based on the *Mean Average Error* (MAE), the absolute difference between ground truth annotation and network predictions, reported in Euler angles.

Method	Data	Pitch	Roll	Yaw
Saeed et al. [3]	RGB+depth	5.0 ± 5.8	4.3 ± 4.6	3.9 ± 4.2
Fanelli et al. [1]	depth	8.5 ± 9.9	7.9 ± 8.3	8.9 ± 13.0
Yang et al. [4]	RGB+depth	9.1 ± 7.4	7.4 ± 4.9	8.9 ± 8.2
Baltruvsaitis et al. [2]	RGB+depth	5.1	11.2	6.29
Papazov et al. [5]	depth	3.0 ± 9.6	2.5 ± 7.4	3.8 ± 16.0
Venturelli et al. [6]	depth	2.8 ± 3.1	2.3 ± 2.9	3.6 ± 4.1
Our	depth	**2.0 ± 1.9**	**2.1 ± 2.0**	**2.3 ± 2.0**

Table 1. Results on *Biwi* Dataset

Method	Data	Pitch	Roll	Yaw
Saeed et al. [3]	RGB+depth	4.9 ± 5.3	4.4 ± 4.6	**5.1 ± 5.4**
Fanelli et al. [1]	depth	5.9 ± 6.3	-	6.3 ± 6.9
Baltruvsaitis et al. [2]	RGB+depth	7.06	10.48	6.90
Our	depth	**4.9 ± 4.6**	**4.2 ± 4.3**	7.5 ± 6.3

Table 2. Results on *ICT-3DHP* database

In real situations, driver images are usually affected by **occlusions**, caused by hand movements or objects like smartphones, garments and similar.

Figure 2. Examples of simulated occlusions

Type of occlusion	Head		
	Pitch	Roll	Yaw
center	11.9 ± 4.8	3.0 ± 3.4	11.0 ±8.2
bottom	8.6 ± 3.8	3.7 ± 2.9	5.0 ±4.2
right	2.5 ± 2.0	2.9 ± 2.8	8.7 ±6.0
top	34.0 ± 18.8	8.4 ± 8.6	8.9 ±6.2
left	2.8 ± 3.1	4.5 ± 3.0	5.5 ±6.1
random	7.8 ± 7.4	3.1 ± 3.5	5.7 ±4.6

Table 3. Results on *Biwi* dataset with simulated occlusions

Embedded Implementation

The proposed system has been developed and tested using an embedded system. The reasons underlying this implementation are varied:

- **Powerful** and **cheap** GPU boards
- **Plug-in** approach
- **Real time** performance
- **Low** energy consumption

Two embedded NVIDIA boards are exploited:

- **NVIDIA Jetson TK1**
  - Cuda cores: 192 *Kepler*
  - CPU: ARM A15
  - RAM: 2 GB

- **NVIDIA Jetson TX1**
  - Cuda cores: 256 *Maxwell*
  - CPU: ARM A57
  - RAM: 4 GB

As **depth** sensor we use:

- **Microsoft Kinect One**
  - Time-of-light
  - RGB: 1920x1080
  - Depth: 512x424
  - 30 fps

Two logical parts have to be implemented: the **acquisition** process, to acquire depth frame, and the **prediction** process, demanded to the RNN described above.
Both parts have been developed in *Python*, exploiting the *Keras* framework with *Theano* back-end.

Boards	Jetson TK1		Jetson TX1	
	CPU	GPU	CPU	GPU
Acquisition Time	0.0305	0.0277	0.0141	0.0238
Prediction Time	0.0717	0.0256	0.0525	0.0091
Total Time	0.1022	**0.0533**	0.0666	**0.0329**

Table 4. Speed performance on Jetson TK1 and TX1

→ **19 fps** on TK1 and **30 fps** on TX1

Acknowlegments

This work has been carried out within the projects "Citta educante" (CTN01-00034-393801) of the National Technological Cluster on Smart Communities funded by MIUR and "FAR2015 - Monitoring the car driver's 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.

References

[1] Fanelli, Gabriele, Juergen Gall, and Luc Van Gool. "Real time head pose estimation with random regression forests." *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.
[2] Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency. "3D constrained local model for rigid and non-rigid facial tracking." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.
[3] Saeed, Anwar and Ayoub Al-Hamadi. "Boosted human head pose estimation using kinect camera." *2015 IEEE International Conference on Image Processing (ICIP)* (2015): 1752-1756.
[4] Yang, Jiaolong, Wei Liang, and Yunde Jia. "Face pose estimation with combined 2D and 3D HOG features." *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012.
[5] Papazov, Chavdar, Tim K. Marks, and Michael Jones. "Real-time 3D head pose and facial landmark estimation from depth images using triangular surface patch features." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
[6] Venturelli Marco, Borghi Guido, Vezzani Roberto, Cucchiara Rita "Deep Head Pose Estimation from Depth Data for In-car Automotive Applications" Workshop on Understanding Human Activities through 3D Sensors, ICPRW, 2016