POSEidon: Face-from-Depth for Driver Pose Estimation

Guido Borghi Marco Venturelli Roberto Vezzani Rita Cucchiara
University of Modena and Reggio Emilia
{name.surname}@unimore.it

Abstract

Fast and accurate upper-body and head pose estimation is a key task for automatic monitoring of driver attention, a challenging context characterized by severe illumination changes, occlusions and extreme poses. In this work, we present a new deep learning framework for head localization and pose estimation on depth images. The core of the proposal is a regressive neural network, called POSEidon, which is composed of three independent convolutional nets followed by a fusion layer, specially conceived for understanding the pose by depth. In addition, to recover the intrinsic value of face appearance for understanding head position and orientation, we propose a new Face-from-Depth model for learning image faces from depth. Results in face reconstruction are qualitatively impressive. We test the proposed framework on two public datasets, namely Biwi Kinect Head Pose and ICT-3DHP, and on Pandora, a new challenging dataset mainly inspired by the automotive setup. Results show that our method overcomes all recent state-of-art works, running in real time at more than 30 frames per second.

1. Introduction

Nowadays, we are witnessing a revolution in the automotive field, where ICT technologies are becoming sometimes more important than the engine itself.

New solutions are required to solve many human-centered problems: semi-autonomous driving, driver behavior understanding, human-machine-interaction for entertainment, driver attention analysis for safe driving are just some examples. All of them, lay on the basic task of estimating driver pose, and in particular of the face and upper body parts, which are the mainly visible items of a driver. Computer vision research [37, 50, 5, 16, 13] achieved encouraging results, even if they are still not completely satisfactory due to some strong constraints of the context: reliability with strong pose changes, robustness to large occlusions (*e.g.* glasses), in conjunction with non-intrusive capabilities, real time and low cost requirements (Fig. 1). In addition, standard techniques based on intensity images are not always applicable, due to the poor illumination conditions during the night and the continuous illumination changes during the day. For this reasons, computer vision solutions based on illumination-insensitive data sources such as thermal [51] or depth [35] cameras are emerging.

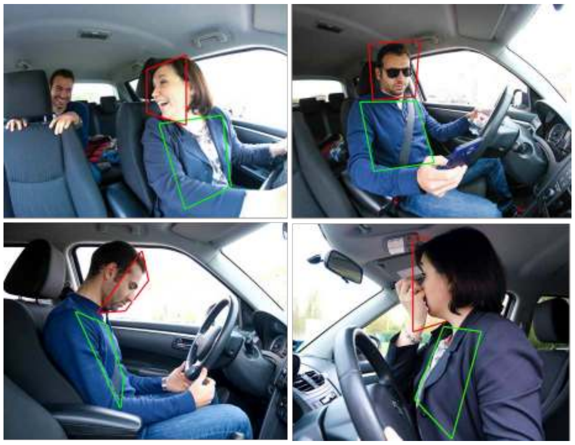
Figure 1. Some real situations in which head and upper-body pose estimation are useful to monitor driver's attention level: from the top-left, driver is talking with passengers, is playing with smartphone, is falling sleep and is looking at the rear-view mirror.

Therefore, we propose a complete framework for driver monitoring based on depth images only, that can be easily acquired by commercial low-cost sensors placed inside the vehicles. Starting from head localization, the ultimate goal of the framework is the estimation of the head and shoulder pose, measured as *pitch*, *roll* and *yaw* rotation angles. To this aim, a new triple regressive Convolutional Neural Network architecture, called *POSEidon*, is proposed, that combines depth, motion images and appearance.

One of the most innovative contribution is a *Face-from-Depth* network, that is able to reconstruct gray-level faces directly from head depth images. This solution derives from the awareness that intensity face images are very useful to detect head pose [1, 17]: without having intensity data we would like to have similar benefits. Gray-level faces extracted by depth images have a qualitatively impressive sim-

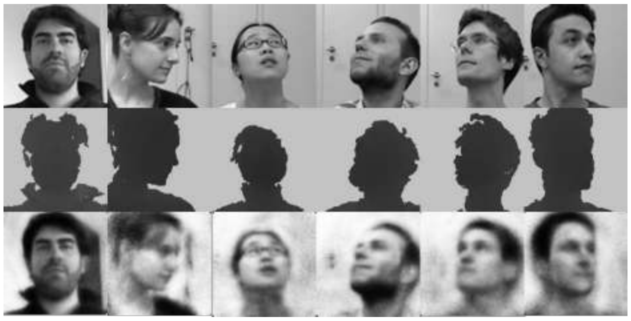

Figure 2. Examples of gray-level face images (bottom) reconstructed from the depth maps (middle). The corresponding ground truth is also shown (top). The first four subjects have been included in the training set, while the last two are completely new.

ilarity (Fig. 2). Summarizing, the novel contributions of the paper are the following:

1. A complete and accurate framework, from head localization to head and shoulder pose estimation, based only on depth data, working in real time (30 fps);

2. A new *Faces-from-Depth* architecture, to reconstruct gray-level face images directly from depth maps. To the best of our knowledge, this is the first proposal of this kind of approach;

3. A new dataset, called *Pandora*, the first containing high resolution depth data with head and shoulder pose annotations.

2. Related Work

Head pose estimation approaches can rely on different input types: intensity images, depth maps, or both. In order to discuss related work, we adopt the classification proposed in [35, 19], updated and summarized in three main categories, namely *feature-based*, *appearance-based* and *3D model registration* approaches.

Feature-based methods need facial (*e.g.* nose, eyes) or pose-dependent features, that should be visible in all poses: consequently, these methods fail when features are not detected. In [31] an accurate nose localization is used for head tracking and pose estimation on depth data. Breitenstein *et al.* [8] used geometric features to identify nose candidates to produce the final pose estimation. HOG features [14] were extracted from RGB and depth images in [55, 44], then a Multi Layer Perceptron and a linear SVM were used for feature classification, respectively. Also [53, 56, 34] needed well visible facial features on RGB input images, and [48] on 3D data.

Appearance-based methods rely on one or more classifiers that use raw input images, trained to perform head pose estimation. In [46] RGB and depth data were combined, exploiting a neural network to perform head pose prediction. Fanelli *et al.* [19, 20, 18] trained Random Regression Forest for both head detection and pose estimation on depth images. Tulyakov *et al.* [52] used a cascade of tree classifiers to tackle extreme head pose estimation task. A Convolutional Neural Network (CNN) based on RGB input images is exploited in [1]. Recently, in [36] a multimodal CNN was proposed to estimate gaze direction: a regression approach was only approximated through a 360-classes classifier. Synthetic datasets were used to train CNNs, that generally require a huge amount of data, *e.g.* [28].

3D model registration approaches create a head model from the acquired data; frequently, a manual initialization is required. In [41] facial point clouds were matched with pose candidates, through a triangular surface patch descriptor. In [3] intensity and depth data were used to build a 3D constrained local method for robust facial feature tracking. In [23] a 3D morphable model is fitted, using both RGB and depth data to predict head pose. Also [3, 9, 6, 10, 7, 43] built 3D facial model for head tracking, animation and pose estimation.

Remaining methods regard head pose estimation task as an *optimization* problem: [39] used the Particle Swarm Optimization (PSO) [25]; [4] exploited the Iterative Closest Point algorithm (ICP) [30]; [35] combined PSO and ICP techniques. [26] used a least-square technique to minimize the difference between the input depth change rate and the prediction rate. Besides, other works use linear or nonlinear regression with extremely low resolution images [11]. HOG features and a Gaussian locally-linear mapping model were used in [17]. Finally, recent works produce head pose estimations performing face alignment task [58].

Several works based on head pose estimation do not take in consideration head localization task. To propose a complete head pose estimation framework, it is necessary to perform a head detection, finding the complete head or a particular point, for example the head center. With RGB images Viola and Jones [54] face detector is often exploited, *e.g.* in [23, 9, 43, 3, 46]. A different approach demands the head location to a classifier, *e.g.* [19, 52]. As reported in [35], these approaches suffer due to the lack of generalization capabilities with different acquisition devices.

Only few works in literature tackle the problem of driver body pose estimation focusing only on upper-body part or in automotive context. Ito *et al.* [24] adopting an intrusive approach, placed six marker points on driver body to predict some typical driving operations. A 2D driver body tracking system was proposed in [15], but a manual initialization of the tracking model is strictly required. In [51] a thermal long-wavelength infrared video camera was used to analyze occupant position and posture. In [49] an approach for upper body tracking system using 3D head and hands movements was developed.

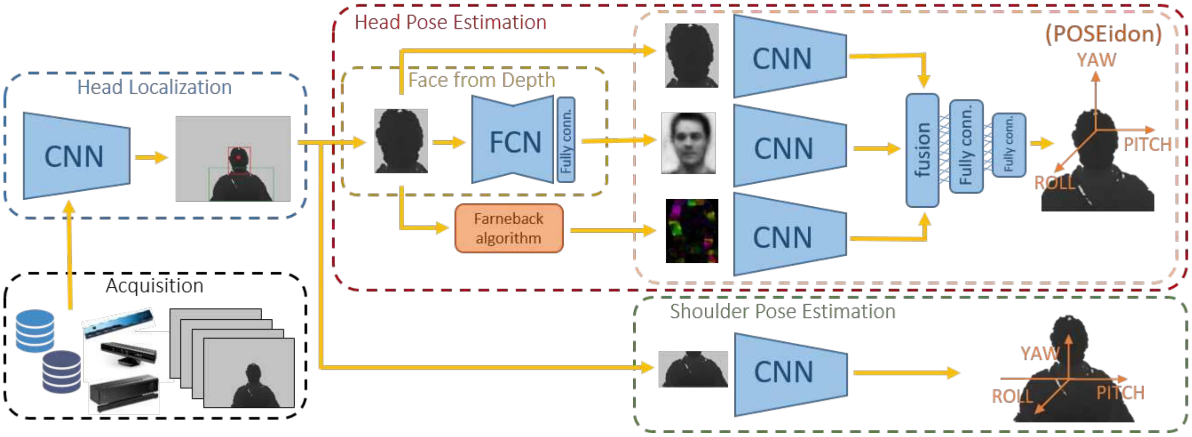

Figure 3. Overview of the whole *POSEidon* framework. Depth input images are acquired by low cost sensors (black) and provided to a head localization CNN (blue) to suitably crop the images around the upper-body or head regions. The first is exploited by the shoulder pose estimation task (green), while the second is selected for the head pose estimation (red) obtained through the *POSEidon* network (orange). In the center, the *Face-from-Depth* net (yellow) which produces gray-level images of the face from the depth map. [best in color]

3. The POSEidon framework

An overview of the *POSEidon* framework is depicted in Figure 3. The final goal is the estimation of the pose of the driver's head and shoulders, defined as the mass center position and the corresponding orientation relatively to the reference frame of the acquisition device [37]. The orientation is represented using three *pitch*, *roll* and *yaw* rotation angles. *POSEidon* directly processes the stream of depth frames captured in real time by a commercial sensor (*e.g.*, *Microsoft Kinect*). Position and size of the head in the foreground are estimated by a head localization module based on a regressive CNN (Sect. 5.1). The output provided is used to crop the input frames around the head or the shoulder bounding boxes, depending on the further pipeline type. Frames cropped around the head are fed to the head pose estimation block, while the others are exploited to estimate the shoulders pose. The core components of the system are the *Face-from-Depth* network (Sect. 4), and *POSEidon* (Sect. 5.2), the network which gives the name to the whole framework. Its trident shape is due to the three included CNNs, each working on likewise sources: depth, *Face-from-Depth* and motion images data. The first one – *i.e.*, the CNN directly working on depth data – plays the main role on the pose estimation, while the other two cooperate to reduce the estimation error.

4. Face-from-Depth network

Face-from-Depth (FfD) is one of the most innovative elements of the framework. Due to illumination issues, the appearance of the face is not always available in many scenarios, *e.g.* inside a vehicle. On the contrary, depth maps are invariant to illumination conditions but lacks of texture details. We aim to investigating if it is possible to imagine the appearance of a face given the corresponding depth data. The *Face-from-Depth* network has been created to this goal, even if the output is not always realistic and visually pleasant: however, the promising results confirm their positive contribution in the estimation of the head pose.
The proposed architecture fuses the key aspects of autoencoders [33] and fully convolutional [29] neural networks: it is composed by 14 convolutional layers, plus a fully connected layer at the end (Fig. 4). A single 2×2 max-pooling layer has been inserted after the second layer, and a corresponding up-sampling layer after the thirteenth. Besides, two zero-padding layers are added after the first and the second convolutional layers, respectively. We train the network in a single stage, with input head images resized to 64×64 pixels. The hyperbolic tangent activation function is used and best training performances are reached through the self adaptive *Adadelta* optimizer [57]. A specific loss function is exploited to highlight the central area of the image, where the face is supposed to be after the cropping step, and takes in account the distance between the reconstructed image and the corresponding gray-level ground truth:

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

where R, C are the number of rows and columns of the input images, respectively. $y_{ij}, \bar{y}_{ij} \in \mathcal{R}^{ch}$ are the intensity values from ground truth ($ch = 1$) and predicted appearance images. Finally, the term $w_{ij}^{\mathcal{N}}$ introduces a bivariate Gaussian prior mask. Best results have been obtained using $\mu = [\frac{R}{2}, \frac{C}{2}]^T$ and $\Sigma = \mathbb{I} \cdot [(R/\alpha)^2, (C/\beta)^2]^T$ with α and β empirically set to $3.5, 2.5$ for squared images of $R = C = 64$. Some visual examples of input, output and ground-truth images are reported in Figure 2.

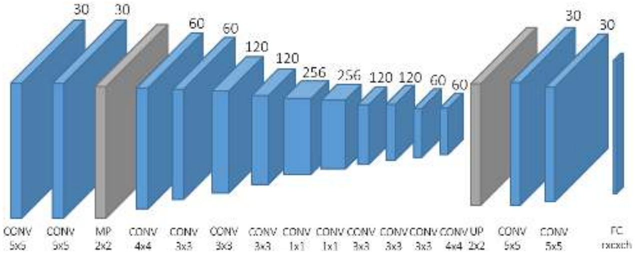

Figure 4. Architecture of the *Face-from-Depth* network.

5. Pose Estimation from depth

5.1. Head Localization Network

In this step we design a network to perform head localization, relying on the main assumption that a single person is in the foreground. The desired network outputs are the image coordinates (x_H, y_H) of the head center, or rather, the average position of all head points in the frame [47]. Details on the deep architecture adopted are reported in Figure 5. A limited depth and small sized filters have been chosen to meet real time constraint while keeping satisfactory performance. For the same reason, input images are firstly resized to 160×132 pixels. A max-pooling layer is run after each of the first four convolutional layers, while a dropout regularization ($\sigma = 0.5$) is exploited in fully connected layers. The hyperbolic tangent activation (*tanh*) function is adopted, in order to map continuous output values to a predefined range $[-\infty, +\infty] \rightarrow [-1, +1]$. The network has been trained by *Stochastic Gradient Descent* (SGD) [27] and the L_2 loss function.

Given the head position (x_H, y_H) in the frame, a dynamic size algorithm provides the head bounding box with barycenter (x_H, y_H) and width w_H and height h_H, around which the frames are cropped:

$$w_H = \frac{f_x \cdot R_x}{D}, \quad h_H = \frac{f_y \cdot R_y}{D} \tag{2}$$

where f_x, f_y are the horizontal and the vertical focal lengths in pixels of the acquisition device, respectively. R_x, R_y are the average width and height of a face (for head pose task $R_x = R_y = 320$) and D is the distance between the head center and the acquisition device, computed averaging the depth values around the head center.

Some examples of the bounding boxes estimated by the network are superimposed in Figure 9.

5.2. POSEidon

The *POSEidon* network is mainly obtained as a fusion of three CNNs and has been developed to perform a regression on the 3D pose angles. As a result, continuous Euler values – corresponding to the *yaw*, *pitch* and *roll* angles – are estimated (right part of Fig. 3). The three *POSEidon* components have the same shallow architecture based on 5 convolutional layers with kernel size of 5×5, 4×4 and 3×3, max-pooling is conducted only on the first three layers. The first four convolutional layers have 32 filters each, the last one has 128 filters. At the end of the network, there are 3 fully connected layers, with 128, 84 and 3 neurons, respectively. Also in this case *tanh* function is exploited: we are aware that *ReLU* [38] converges faster, but we obtain better performance in term of accuracy prediction. The three networks are fed with different input data types: the first one, directly takes as input the head-cropped depth images; the second one is connected to the *Face-from-Depth* output and the last one operates on motion images, obtained applying the standard *Farneback* algorithm [21] on pairs of consecutive depth frames. A fusion step combines the contributions of the three above described networks: in this case, the last fully connected layer of each component is removed. Different fusion approaches that have been proposed [42] are investigated. Given two feature maps x^a, x^b with a certain width w and height h, for every feature channel d^x_a, d^x_b and $y \in R^{w \times h \times d}$:

- **Multiplication**: computes the element-wise product of two feature maps, as $y^{mul} = x^a \circ x^b, d^y = d^x_a = d^x_b$
- **Concatenation**: stacks two features maps, without any blend $y^{cat} = [x^a | x^b], d^y = d^x_a + d^x_b$
- **Convolution**: stacks and convolves feature maps with a filter k of size $1 \times 1 \times (d^x_a + d^x_b)/2$ and β as bias term, $y^{conv} = y^{cat} * k + \beta, \quad d^y = (d^x_a + d^x_b)/2$

Final *POSEidon* framework exploits a combination of two fusing methods, in particular a convolution followed by a concatenation. After the fusion step, three fully connected layers composed of 128, 84 and 3 activations respectively and two dropout regularization ($\sigma = 0.5$) complete the architecture. *POSEidon* is trained with a double-step procedure. First, each individual network is trained with the fol-

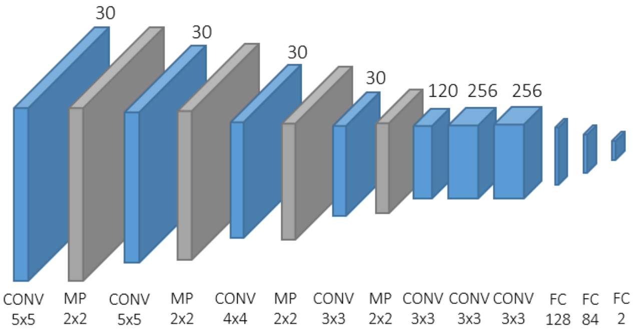

Figure 5. Architecture of the head localization network.

lowing L_2^w weighted loss:

$$L_2^w = \sum_{i=1}^{3} \left\| w_i \cdot (y_i - f(x_i)) \right\|_2 \quad (3)$$

where $w_i \in [0.2, 0.35, 0.45]$: this weight distribution gives more importance to the yaw angle, which is preponderant in the selected automotive context. During the individual training step, the last fully connected layer of each network is preserved, then is removed to perform the second training phase: holding the weights learned for the trident components, the new training phase is carried out on the last three fully connected layers of *POSEidon*, with the loss function L_2^w reported in Equation 3. In all training steps, the SGD optimizer [27] is exploited, the learning rate is set initially to 10^{-1} and then is reduced by a factor 2 every 15 epochs.

5.3. Shoulder pose estimation network

The framework is completed with an additional network for the estimation of the shoulder pose. We employ the same architecture adopted for the head (Section 5.2), performing regression on the same three pose angles. Starting from the head center position (Section 5.1), the depth input images are crop around the driver neck, using a bounding box $\{x_S, y_S, w_S, h_S\}$ with barycenter $(x_S = x_H, y_S = y_H - (h_H/4))$, and width and height obtained as described in Equation 2, but with different values of R_x, R_y to produce a rectangular crop: these values are tested and discussed in Section 7. The network is trained with SGD optimizer [27], using the weighted L_2^w loss function described above (see Eq. 3). Hyperbolic tangent is exploited as activation functions as usual.

6. Datasets

Network training and testing phases have been done exploiting two publicly available datasets, namely *Biwi Kinect Head Pose* and *ICT-3DHP*. In addition, we collect a new dataset, called *Pandora*, which also contains shoulder pose annotations. Data augmentation techniques are employed to enlarge the training set, in order to achieve space invariance and avoid over fitting [27]. Random translations on vertical, horizontal and diagonal directions, jittering, zoom-in and zoom-out transformation of the original images have been exploited. Percentile-based contrast stretching, normalization and scaling of the input images are also applied to produce zero mean and unit variance data.
Follows a detailed description of the three adopted datasets.

6.1. Biwi Kinect Head Pose dataset

Fanelli *et al.* [19] introduced this dataset in 2013. It is acquired with the *Microsoft Kinect* sensor, a structured IR light device. It contains about 15k frame, with RGB (640 × 480) and depth maps (640 × 480). Twenty subjects have been involved in the recordings: four of them were recorded twice, for a total of 24 sequences. The ground truth of yaw, pitch and roll angles is reported together with the head center and the calibration matrix. The original paper does not report the adopted split between training and testing sets; fair comparisons are thus not guarantee. To avoid this, we clearly report the adopted split in the following.

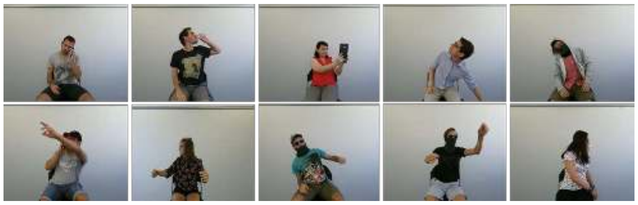

Figure 6. Sample frames from the *Pandora* dataset.

6.2. ICT-3DHP dataset

ICT-3DHP dataset has been introduced by Baltrusaitis *et al.* in 2012 [3]. It is collected using the *Microsoft Kinect* sensor and contains RGB images and depth maps of about 14k frames, divided in 10 sequences. The image resolution is 640 × 480 pixels. An hardware sensor (*Polhemus Fastrack*) is exploited to generate the ground truth annotation. The device is placed on a white cap worn by each subject, visible in both RGB and depth frames. Moreover, the presence of few subjects and the limited number of frames make this dataset unsuitable for deep learning approaches.

6.3. Pandora dataset

We collect a new challenging dataset, called *Pandora*. The dataset has been specifically created for the tasks described in the paper (*i.e.*, head center localization, head pose and shoulder pose estimation) and is inspired by the automotive context. A frontal fixed device acquires the upper body part of the subjects, simulating the point of view of camera placed inside the dashboard. Among the others, the subjects also perform driving-like actions, such as grasping the steering wheel, looking to the rear-view or lateral mirrors, shifting gears and so on. *Pandora* contains 110 annotated sequences using 10 male and 12 female actors. Each subject has been recorded five times.

Pandora is the first publicly available dataset which combines the following features:

- **Shoulder angles**: in addition to the head pose annotation, *Pandora* contains the ground truth data of the shoulder pose expressed as yaw, pitch and roll.

- **Wide angle ranges**: subjects perform wide head (±70° roll, ±100° pitch and ±125° yaw) and shoulder (±70° roll, ±60° pitch and ±60° yaw) movements.

HEAD POSE ESTIMATION ERROR [EULER ANGLES]

Method	Year	Data	Pitch	Roll	Yaw	Avg
Fanelli [19]	2011	Depth	8.5 ± 9.9	7.9 ± 8.3	8.9 ± 13.0	8.43 ± 10.4
Yang [55]	2012	RGB + Depth	9.1 ± 7.4	7.4 ± 4.9	8.9 ± 8.3	8.5 ± 6.9
Padeleris [39]	2012	Depth	6.6	6.7	11.1	8.1
Rekik [43]	2013	RGB + Depth	4.3	5.2	5.1	4.9
Baltrusaitis [3]	2012	RGB + Depth	5.1	11.3	6.3	7.6
Ahn [1]*	2014	RGB	3.4 ± 2.9	2.6 ± 2.5	2.8 ± 2.4	2.9 ± 2.6
Martin [32]*	2014	Depth	2.5	2.6	3.6	2.9
Saeed [44]	2015	RGB + Depth	5.0 ± 5.8	4.3 ± 4.6	3.9 ± 4.2	4.4 ± 4.9
Papazov [41]	2015	Depth	2.5 ± 7.4	3.8 ± 16.0	3.0 ± 9.6	4.0 ± 11.0
Drouard [17]	2015	RGB	5.9 ± 4.8	4.7 ± 4.6	4.9 ± 4.1	5.2 ± 4.5
Meyer [35]	2015	Depth	2.4	2.1	2.1	2.2
Liu [28]	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**

Table 1. Results on *Biwi* dataset. Input cropping is done using the ground truth head position.

For each subject, two sequences are performed with constrained movements, changing the yaw, pitch and roll angles separately, while three additional sequences are completely unconstrained.

- **Challenging camouflage**: garments as well as various objects are worn or used by the subjects to create head and/or shoulder occlusions. For example, people wear prescription glasses, sun glasses, scarves, caps, and manipulate smartphones, tablets or plastic bottles.
- **Deep-learning oriented**: the dataset contains more than 250k full resolution RGB (1920 × 1080) and depth images (512 × 424) with the corresponding annotation.
- **Time-of-Flight (ToF) data**: a *Microsoft Kinect One* device is used to acquire depth data, with a better quality than other datasets created with the first *Kinect* version [45].

Each frame of the dataset is composed of the RGB appearance image, the corresponding depth map, the 3D coordinates of the skeleton joints corresponding to the upper body part, including the head center and the shoulder positions. For convenience's sake, the 2D coordinates of the joints on both color and depth frames are provided as well as the head and shoulder pose angles with respect to the camera reference frame. Shoulder angles are obtained through the conversion to Euler angles of a corresponding rotation matrix, obtained from a user-centered system [40] and defined by the following unit vectors (N_1, N_2, N_3):

$$
\begin{aligned}
N_1 &= \frac{p_{RS}-p_{LS}}{\|p_{RS}-p_{LS}\|} \quad & U &= \frac{p_{RS}-p_{SB}}{\|p_{RS}-p_{SB}\|} \\
N_3 &= \frac{N_1 \times U}{\|N_1 \times U\|} \quad & N_2 &= N_1 \times N_3
\end{aligned}
\tag{4}
$$

where p_{LS}, p_{RS} and p_{SB} are the 3D coordinates of the left shoulder, right shoulder and spine base joints. The annotation of the head pose angles has been collected using a wearable *Inertial Measurement Unit* (IMU) sensor. The sensor has been worn by the subjects in a non visible position (*i.e.*, on the rear of the head) to avoid distracting artifacts on both color and depth images. IMU sensor has been calibrated and aligned at the beginning of each sequence, assuring the reliability of the provided angles. The dataset is publicly available (http://imagelab.ing.unimore.it/pandora/).

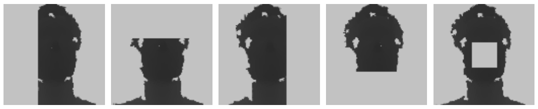
(a) left (b) top (c) right (d) bottom (e) middle

Figure 7. Visual examples of the simulated occlusion types applied on a *Biwi* frame.

7. Experimental results

The proposed framework has been deeply tested using dataset described in Section 6. Besides, an ablation study has been evaluated on *Pandora*.

Sequences of subjects 10, 14, 16 and 20 have been used for testing, the remaining for training. Table 2 reports an internal evaluation, providing mean and standard deviation of the estimation errors obtained on each angle and for each system configuration. Similar to Fanelli *et al.* [19], we also report the mean accuracy as percentage of good estimations (*i.e.*, angle error below 15°). The first line of Table 2 shows the performance of a baseline system, obtained using the pose estimation network only and input depth frames are directly fed to the network without processing and crop. The crop step is included instead in the configurations of the other rows, using the ground truth head position as center. Results obtained using single networks, couples of them and the complete *POSEidon* architecture are shown. The

HEAD POSE ESTIMATION ERROR [EULER ANGLES]

Architecture	Input	Cropping	Fusion	Head			Accuracy
				Pitch	Roll	Yaw	
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

Table 2. Results of the head pose estimation on *Pandora* comparing different system architectures. The baseline is a single CNN working on the source depth map. The accuracy is the percentage of correct estimations (err <15°). FfD: Face-from-Depth, MI: Motion Images.

last row highlights the best performance reached using *conv* fusion of couples of input types, followed by the *concat* step. Even if the choice of the fusion method has a limited effect (as deeply investigated in [42, 22]), the most significant improvement of the system is reached exploiting the three input types together.

Figure 8 shows a comparison of the estimation errors made by each trident component: each graph plots the error distribution of a specific network configuration with respect to the ground truth value. Depth data allows to reach the lowest error rates for frontal heads, while the other input data types are better in presence of rotated poses. The graphs highlight the averaging capabilities of *POSEidon* too.

Table 2 includes an indirect evaluation of the reconstruction capabilities of the *Face-from-Depth* network. The results reported on the third and fourth rows are obtained using the network described in Section 5.2 with the reconstructed appearance image and the original gray-level images as input, respectively. The similar results confirm that the obtained image reconstruction is sufficiently accurate, at least for the pose estimation task. We compared the results of *POSEidon* with state-of-art, using the *Biwi* dataset. According with Fanelli *et al.* [19], 18 subjects are used to train the system while two for the test. More specifically, we exploited the sequences 11 and 12 for testing and the remaining for training. Table 1 reports the corresponding results as indicated in the cited papers. *POSEidon* achieves impressive results on *Biwi* dataset: the mean error is under 2° for all of the three angles, with a small standard deviation. The system overcomes all the reported methods, included the recent proposal by Meyer *et al.* [35]. The performance are better than other approaches based on deep learning, 3D data and regression [1, 36]. Moreover, *POSEidon* also overcomes the approaches working on appearance data. The proposals marked with a star (*) do not follow the same split or apply a different testing procedure: thus, the comparison with them may not be fair. Results of [39] reported in table have been taken from [35] for the sake of comparability.

As already mentioned, in real situations the driver head may be affected by severe occlusions caused by hands and objects such as smartphones, scarves, bottles and so on. For this reason, we have carried out a specific set of experiments to test the reliability of *POSEidon* in presence of occlusions or missing data. We artificially applied the masks depicted in Figure 7 to remove parts of the input frames and simulate occlusions. The corresponding performance of *POSEidon* is shown in Table 3, which confirms the reliability of the system also in these cases. The absence of the upper body part of the head strongly impacts with the system performance, in particular for the estimation of the pitch angle. Similarly, the head part around the nose plays a crucial role in the pose estimation, as highlighted by the errors gener-

Occluded part	Head		
	Pitch	Roll	Yaw
(a) left	2.6± 3.0	4.0 ± 2.9	7.8 ± 8.1
(b) top	42.5 ± 21.2	12.3 ± 9.3	10.2 ± 7.6
(c) right	2.1 ± 1.8	2.8 ± 2.6	8.4 ± 8.5
(d) bottom	4.2 ± 3.3	4.3 ± 3.5	4.0 ± 3.0
(e) middle	11.0 ± 5.3	3.0 ± 2.8	6.1 ± 4.9
random	12.5 ± 18.3	5.3 ± 6.1	7.4 ± 7.1

Table 3. Estimation errors of *POSEidon* in presence of simulated occlusions. The system is fed with images from the *Biwi* dataset occluded using the masks illustrated in figure 7. Results of the last line are obtained by applying a random mask to each frame.

Parameters		Shoulders			Accuracy
R_x	R_y	Pitch	Roll	Yaw	
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**

Table 4. Estimation errors and mean accuracy of the shoulder pose estimation on *Pandora*

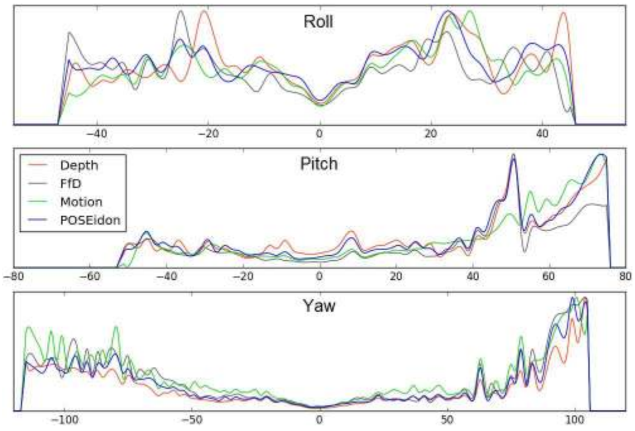

Figure 8. Error distribution of each *POSEidon* components on *Pandora* dataset. On x-axis are reported the ground truth angles, on y-axis the distribution of error for each input type.

ated by the occlusion type (e).

The network performing the shoulder pose estimation has been tested on *Pandora* only, due to the lack of the corresponding annotations in the other datasets. Results are reported in Table 4, where different image crops are compared (Section 5.3). The reported results are very promising, reaching an accuracy over 92%.

In order to have a fair comparison, results reported in Tables 1 and 2 are obtained using the ground truth head position as input to the crop procedure. We finally test the whole pipeline, including the head localization network described in section 5.1, using also *ICT-3DHP* dataset. The mean error of the head localization (in pixels) and the pose estimation errors are summarized in Table 5. Sometimes, the estimated position generates a more effective crop of the head. As a result, the whole pipeline performs better on the head pose estimation over the *Biwi* dataset. *POSEidon* reaches valuable results also on the *ICT-3DHP* dataset and it provides comparable results with respect to state of the art methods working on both depth and RGB data (4.9±5.3, 4.4±4.6, 5.1±5.4 [44], 7.06, 10.48, 6.90 [3], for pitch, roll and yaw respectively).

The complete framework has been implemented and tested on a desktop computer equipped with a *NVidia Quadro k2200* GPU board and on a laptop with a *NVidia GTX 860M*, exploiting *Keras* [12] with *Theano* [2] backend. Real time performance has been obtained in both cases, with a processing rate of more than 30 frames per second, with a limited dedicated graphical memory requirement. Some examples of the system output are reported in Figure 9, where the six pose angles are visually shown using colored bars. In addition, the original depth map, the *Face-from-Depth* reconstruction and the motion data given in input to *POSEidon* are placed on the left of each frame. Pre-trained networks and models are publicly available.

Dataset	**Loc.**	**Head**		
		Pitch	Roll	Yaw
Biwi	3.27±2.19	1.5±1.4	1.7±1.7	2.3±2.1
ICT-3DHP	-	5.0±4.3	3.5±3.5	7.1±6.1
Pandora	4.27±3.25	7.6±8.5	4.8±4.8	10.6±12.7

Table 5. Results on *Biwi*, *ICT-3DHP* and *Pandora* dataset of the complete *POSEidon* pipeline (*i.e.*, head localization, cropping and pose estimation).

8. Conclusions and future work

A complete framework for head localization and driver pose estimation called *POSEidon* is presented. No previous computation of specific facial features is required. The system has shown real time and impressive results also in presence of occlusions, extreme poses of head and shoulders. Besides, the use of only depth data enhances the efficacy under different illumination conditions. All these aspects make the proposed framework suitable to particular challenging contexts, such as automotive. A new and high quality 3D dataset, *Pandora*, is then proposed and publicly released. The system has been developed with a modular architecture: if it is possible to capture both RGB and depth images during the training, the complete architecture can be used. Otherwise, the *Face-from-Depth* module should be removed from the system, using the depth+MI combination, reaching worst but still satisfactory performances.

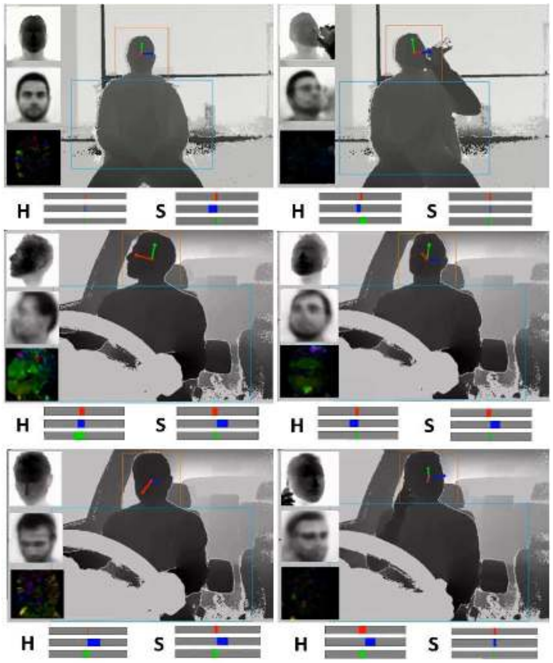

Figure 9. Visual examples of the proposed framework output. Head (H) and shoulder (S) pose angles are reported as bars centered at 0°. Depth maps, *Face-from-Depth* and motion image inputs are depicted on the left of each frame. [best in colors]

References

[1] B. Ahn, J. Park, and I. S. Kweon. Real-time head orientation from a monocular camera using deep neural network. pages 82–96, 2014.

[2] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, et al. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016.

[3] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2610–2617, 2012.

[4] T. Bär, J. F. Reuter, and J. M. Zöllner. Driver head pose and gaze estimation based on multi-template icp 3-d point cloud alignment. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pages 1797–1802, 2012.

[5] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez. Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):63–77, 2006.

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[7] A. Bleiweiss and M. Werman. Robust head pose estimation by fusing time-of-flight depth and color. In *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 116–121, 2010.

[8] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister. Real-time face pose estimation from single range images. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[9] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang. 3d deformable face tracking with a commodity depth camera. In *Proc. of European Conference on Computer Vision*, pages 229–242, 2010.

[10] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013.

[11] J. Chen, J. Wu, K. Richter, J. Konrad, and P. Ishwar. Estimating head pose orientation using extremely low resolution images. In *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 65–68, 2016.

[12] F. Chollet. keras. https://github.com/fchollet/keras, 2015.

[13] B. Czupryński and A. Strupczewski. High accuracy head pose tracking survey. In *International Conference on Active Media Technology*, pages 407–420. Springer, 2014.

[14] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[15] A. Datta, Y. Sheikh, and T. Kanade. Linear motion estimation for systems of articulated planes. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[16] A. Doshi and M. M. Trivedi. Head and eye gaze dynamics during visual attention shifts in complex environments. *Journal of vision*, 12(2):9–9, 2012.

[17] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, and R. Horaud. Head pose estimation via probabilistic high-dimensional regression. In *Proc. of IEEE International Conference on Image Processing*, pages 4624–4628, 2015.

[18] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *Int. J. Comput. Vision*, 101(3):437–458, 2013.

[19] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 617–624, 2011.

[20] G. Fanelli, T. Weise, J. Gall, and L. Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint Pattern Recognition Symposium*, pages 101–110, 2011.

[21] G. Farneback. Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field. In *Proc. of IEEE International Conference on Computer Vision*, volume 1, pages 171–177. IEEE, 2001.

[22] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. *arXiv preprint arXiv:1604.06573*, 2016.

[23] R. S. Ghiass, O. Arandjelović, and D. Laurendeau. Highly accurate and fully automatic head pose estimation from a low quality consumer-level rgb-d sensor. In *Proc. of the 2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, pages 25–34, 2015.

[24] T. Ito and T. Kanade. Predicting driver operations inside vehicles. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[25] J. Kennedy. Particle swarm optimization. In *Encyclopedia of machine learning*, pages 760–766. Springer, 2011.

[26] F. A. Kondori, S. Yousefi, H. Li, S. Sonning, and S. Sonning. 3d head pose estimation using the kinect. In *Proc. of International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–4, 2011.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei. 3d head pose estimation with convolutional neural network trained on synthetic images. In *Proc. of IEEE International Conference on Image Processing*, pages 1289–1293, 2016.

[29] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[30] K.-L. Low. Linear least-squares optimization for point-to-plane icp surface registration. *Techrep - Chapel Hill, University of North Carolina*, 4, 2004.

[31] S. Malassiotis and M. G. Strintzis. Robust real-time 3d head pose estimation from range data. *Pattern Recognition*, 38(8):1153–1165, 2005.

[32] M. Martin, F. v. d. Camp, and R. Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In *Proceedings of the 2014 2Nd International Conference on 3D Vision - Volume 01*, 3DV '14, pages 641–648, Washington, DC, USA, 2014. IEEE Computer Society.

[33] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer, 2011.

[34] Y. Matsumoto and A. Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 499–504, 2000.

[35] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz. Robust model-based 3d head pose estimation. In *Proc. of IEEE International Conference on Computer Vision*, pages 3649–3657, 2015.

[36] S. S. Mukherjee and N. M. Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015.

[37] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, Apr. 2009.

[38] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[39] P. Padeleris, X. Zabulis, and A. A. Argyros. Head pose estimation on depth data based on particle swarm optimization. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 42–49, 2012.

[40] G. T. Papadopoulos, A. Axenopoulos, and P. Daras. Real-time skeleton-tracking-based human action recognition using kinect data. In *International Conference on Multimedia Modeling*, pages 473–483, 2014.

[41] C. Papazov, T. K. Marks, and M. Jones. Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4722–4730, 2015.

[42] E. Park, X. Han, T. L. Berg, and A. C. Berg. Combining multiple sources of knowledge in deep cnns for action recognition. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[43] A. Rekik, A. Ben-Hamadou, and W. Mahdi. 3d face pose tracking using low quality depth cameras. In *VISAPP (2)*, pages 223–228, 2013.

[44] A. Saeed and A. Al-Hamadi. Boosted human head pose estimation using kinect camera. In *Proc. of IEEE International Conference on Image Processing*, pages 1752–1756, 2015.

[45] H. Sarbolandi, D. Lefloch, and A. Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Comput. Vis. Image Und.*, 139:1–20, 2015.

[46] E. Seemann, K. Nickel, and R. Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Proc. of Sixth International Conference on Face and Gesture Recognition*, pages 626–631. IEEE Computer Society, 2004.

[47] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[48] Y. Sun and L. Yin. Automatic pose estimation of 3d facial models. In *Proc. of International Conference on Pattern Recognition*, pages 1–4, 2008.

[49] C. Tran and M. M. Trivedi. Introducing xmob: Extremity movement observation framework for upper body pose tracking in 3d. In *Proc. of IEEE International Symposium on Multimedia*, pages 446–447. IEEE, 2009.

[50] C. Tran and M. M. Trivedi. Vision for driver assistance: Looking at people in a vehicle. In *Visual Analysis of Humans*, pages 597–614. Springer, 2011.

[51] M. M. Trivedi, S. Y. Cheng, E. M. Childers, and S. J. Krotosky. Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. *IEEE Transactions on Vehicular Technology*, 53(6):1698–1712, 2004.

[52] S. Tulyakov, R.-L. Vieriu, S. Semeniuta, and N. Sebe. Robust real-time extreme head pose estimation. In *Proc. of International Conference on Pattern Recognition*, pages 2263–2268, 2014.

[53] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based head pose estimation from images. pages 330–335, 2007.

[54] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[55] J. Yang, W. Liang, and Y. Jia. Face pose estimation with combined 2d and 3d hog features. In *Proc. of International Conference on Pattern Recognition*, pages 2492–2495, 2012.

[56] R. Yang and Z. Zhang. Model-based head pose tracking with stereovision. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 255–260, 2002.

[57] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[58] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face Alignment Across Large Poses: A 3D Solution. *ArXiv e-prints*, Nov. 2015.