Hand Gestures for the Human-Car Interaction: the Briareo dataset

Fabio Manganaro, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara

University of Modena and Reggio Emilia, Italy
Department of Engineering "Enzo Ferrari"
{s.pini, name.surname}@unimore.it

Abstract. *Natural User Interfaces* can be an effective way to reduce driver's inattention during the driving activity. To this end, in this paper we propose a new dataset, called *Briareo*, specifically collected for the hand gesture recognition task in the automotive context. The dataset is acquired from an innovative point of view, exploiting different kinds of cameras, *i.e.* RGB, infrared stereo, and depth, that provide various types of images and 3D hand joints. Moreover, the dataset contains a significant amount of hand gesture samples, performed by several subjects, allowing the use of deep learning-based approaches. Finally, a framework for hand gesture segmentation and classification is presented, exploiting a method introduced to assess the quality of the proposed dataset.

Keywords: Hand Gesture Classification · Automotive Dataset · Driver Attention Monitoring · Deep Learning · C3D · LSTM

1 Introduction

Natural User Interfaces (NUIs), *i.e.* interfaces in which the interaction is not carried through physical devices (like mouses and keyboards), are becoming more and more important in many computer vision fields and a key component of new technological tools, since they are extremely *user-friendly* and *intuitive* [9].
Recently, NUIs are gathering attention also in the *automotive* context, where they can be used for a variety of applications in order to reduce driver inattention. In fact, they can increase the amount of time in which driver attention is focused on driving activity. Indeed, driver distraction, according to the *National Highway Traffic Safety Administration*[1] (NHTSA), is generally defined as "an activity that could divert a person's attention away from the primary task of driving", and is one of the most important causes in fatal road crashes [5].
Generally, three types of driver distraction are identified in the literature [1,2]:

- **Manual Distraction**: driver's hands are not on the steering wheel for a prolonged amount of time. As a consequence, the driver is not ready to avoid road obstacles, such as cars and pedestrians;

[1] https://www.nhtsa.gov

- **Visual Distraction**: driver's eyes are not looking at the road, since they are engaged in different tasks, such as reading a newspaper or looking at the phone;
- **Cognitive Distraction**: driver's attention is not focused on the driving activity due to the *fatigue*, *i.e.* "the inability of disinclination to continue an activity, generally because the activity has been going for too long" [10], or due to bad physical conditions or the cognitive load due to external factors.

The availability of systems that can be controlled via the *Natural Language*, like vocal commands or hand gestures [3], could significantly reduce the causes of manual and visual distraction since they generally lead to a reduction of the amount of time involved in interactive activities. Besides, as reported in [12], today drivers are more engaged in secondary tasks than in the past due to the presence, for instance, of smartphones.

For these reasons, in this paper we investigate the development of a hand gesture-based interaction system, based on computer vision techniques, aiming to obtain a safer interaction between the driver and the car system. A key element in its development is the collection of a new dataset, called *Briareo*, specifically designed for the driver hand gesture classification and segmentation with deep learning-based approaches, which includes a significant amount of annotated samples.
In particular, we focus on *dynamic* hand gestures, *i.e.* each gesture is a combination of motion and one or more hand poses: thus, we neglect static hand gestures, that are out of the scope of this paper. Images have been collected from an innovative point of view, different from other perspectives proposed in the past literature: the acquisition devices are placed in the central tunnel between the driver and the passenger seats, orientated towards the car ceiling. In this way, visual occlusions produced by driver's body can be mitigated.
To collect the dataset, three main requirements about the automotive context have been taken into account [18]:

- **Light Invariance**: vision-based systems have to be reliable even in presence of dramatic light changes (generated, for instance, by the alternation between day and night, tunnels, or bad weather conditions);
- **Non-invasinevess**: driver's movements and gaze must not be impeded during the driving activity. Consequently, sensors have to be easily integrated into the car dashboards;
- **Real Time performance**: interaction systems have to quickly detect gestures and provide a fast feedback of the system;

To tackle the first requirement, we propose the use of infrared-based sensors. Moreover, we select devices that are also able to acquire *depth maps*, *i.e.* particular types of images in which each pixel corresponds to the distance between the acquisition device and that point in the scene. Recently, several infrared and depth devices with high-quality sensors and with a small factor form have been introduced, which fulfil the second requirement.

The rest of the paper is organized as follow. In the next section, related datasets and methods about hand gesture classification are analyzed. Then, in Section 3, the *Briareo* dataset is presented, detailing all the features and data collected. In Section 4, two baseline methods are proposed, in order to assess the quality of the proposed dataset and to move towards the development of a gesture-based interaction framework. Experimental results are presented in Section 5. Finally, Section 6 draws the conclusions.

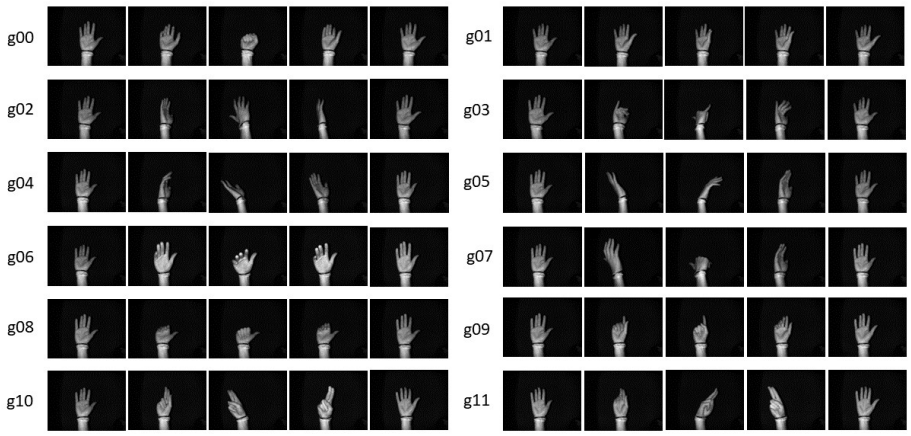

Fig. 1. Gesture classes included in the *Briareo* dataset. As shown, only *dynamic* gestures are present in the dataset. For further details, see Section 3.2.

2 Related Datasets

Recently, several public datasets have been presented in the literature about the driver gesture classification task [11,15,14]. These datasets propose various gesture classes, performed by multiple subjects, with diverse gesture complexity and sensors used for the acquisition part. A summary of these datasets is reported in Table 1.

The *Chalearn* dataset [6] contains a high number of subjects and samples, but it is based only on the *Italian Sign Language* and it is acquired in an indoor environment. The automotive dataset called *Turms* [1] is acquired in a real automotive context, but it is focused on driver's hand detection and tracking, then no hand gestures are present.

The dataset proposed in [11] contains both 3D hand joints information and depth maps, acquired jointly with a *Leap Motion* device and the first version of the *Microsoft Kinect*. There are 10 different gestures performed by 14 people, and each gesture is repeated for 10 times. The acquisition has been conducted in an indoor environment and the devices are frontally placed with respect to the

subjects. Unfortunately, hand gestures are *static* and belong to the *American Sign Language*.

The *VIVA Hand Gesture Dataset* [15] is a dataset released for the namesake challenge, organized by the *Laboratory for Intelligent and Safe Automobiles* (LISA). This dataset has been designed to study natural human activities in confused and difficult contexts, with a variable illumination and frequent occlusions. 19 gesture classes are reported, taken from 8 different subjects, simulating real driving situations. Authors provide both RGB and depth maps acquired using the first version of the *Microsoft Kinect*. It is worth noting that users perform gestures around the infotainment area, placing the right hand on a green and flat surface to facilitate vision-based algorithms. The best gesture recognition method proposed in the challenge consists of a 3D convolutional neural network-based algorithm which has been presented by Molchanov *et al.* in [13].

The *Nvidia Dynamic Hand Gesture* dataset [14] presents 25 types of gestures recorded by multiple sensors (*SoftKinetic DS235* and a *DUO 3D* stereo camera) from different points of view: acquisition devices are frontal placed and top-mounted with respect to the driver position. The acquisition has been carried out in an indoor car simulator. Users perform gestures with the right hand while the left one grasps the steering wheel. The dataset contains the recordings of 20 subjects, even if some of them contributed only partially, not performing the entire recording session. In addition, optical flow is computed on intensity images and it is publicly released.

The *Leap Motion Dynamic Hand Gesture* (LMDHG) dataset [4] contains unsegmented dynamic gestures, performed with either one or two hands. The *Leap Motion* sensor has been employed as acquisition device because its SDK is able to extract the 3D coordinates of 23 hand joints. This dataset is composed of several sequences executed by 21 participants and it contains 13 types of gestures performed randomly alongside an additional no-gesture action. Overall, 50 sequences are released, leading to a total of 608 gesture instances.

Table 1. Datasets for the hand gesture classification task. We report the number of subjects and gesture classes and the types of data included: RGB images, depth maps (acquired with *Structured Light* (SL) or *Time-of-Flight* (ToF) devices), infrared images. Moreover, we report the presence of 3D hand joints (3DJ) and dynamic gestures.

Dataset	**Year**	**#subjs**	**#gest**	**RGB**	**Depth**	**IR**	**3DJ**	**Dynamic**
Unipd [11]	2014	14	10	√	SL		√	
VIVA [15]	2014	8	19	√	SL	√		√
Nvidia [14]	2015	20	25	√	SL	√		√
LMDHG [4]	2017	21	13			√	√	√
Turms [1]	2018	7	-			√		√
Briareo	2019	40	12	√	ToF	√	√	√

3 The Briareo dataset

In this Section, we introduce the *Briareo* dataset, highlighting the original contributions of the proposed data collection with respect to the previous ones.
As mentioned above, this dataset contains *dynamic* hand gestures, shown in Figure 1, acquired indoor in a real car dashboard. Furthermore, the dataset introduces an innovative point of view, not used in previous datasets: we place the acquisition devices in the central tunnel, between the driver and the passenger seat. This choice has been driven by the hypothesis that from this point of view it is possible to acquire gestures with minor visual occlusions compared to other camera positions. Moreover, in this position the acquisition devices can be easily integrated and are protected by direct sunlight, which is a critical element for infrared-based sensors.
Finally, this dataset contains a great variability in the collected data: a high number of subjects and gestures have been recorded. The great amount of annotated data allows using deep learning-based techniques. The dataset is publicly available[2].

Fig. 2. Sample infrared, RGB, and depth images from the *Briareo* dataset. Samples have been acquired with a standard RGB camera and the *Pico Flexx* device.

3.1 Acquisition Devices

Three different sensors are used in order to acquire the dataset.
Firstly, a traditional RGB camera is exploited, able to acquire data up to 30 frames per second. In order to maintain the realism of the automotive environment, no external light sources have been added: this results in dark intensity frames with low contrast, as depicted in Figure 2. Secondly, we used a depth sensors, namely the *Pico Flexx*[3], which has the following features:

– **Time-of-Flight (ToF)**: thanks to this technology, the device is able to acquire 16-bit depth maps with a spatial resolution of 224×171. As reported in [16], ToF technology provides better quality and a faster frame rate than

[2] http://imagelab.ing.unimore.it/briareo
[3] https://pmdtec.com/picofamily/flexx

the *Structured Light* devices (*e.g.* the first version of the *Microsoft Kinect*), reducing the number of visual artefacts, like holes and missing values;
- **Factor Form**: the sensor has very limited dimensions (68mm × 17mm × 7.35mm) and weight (8g), so it can be easily integrated in the car cockpit;
- **High Framerate**: different work modalities are available for the device: selecting a limited acquisition range, it is able to acquire up to 45 frames per second. This is a crucial element in order to achieve real time performance;
- **Acquisition range**: there are two possible depth resolutions, the first one acquires objects in the range 0.5−4 m, while the second one in the range 0.1−1 m. We set the second modality: in this way, the sensor is able to correctly acquire gestures performed close to the device. Indeed, we hypothesize that a distance grater than 1 meter is useless in our acquisition setting.

Finally, we employ an infrared stereo camera, the *Leap Motion*[4], with the following features:

- **Infrared cameras**: the device has two infrared cameras with a resolution of 640 × 240 and 400 × 400 pixels for raw and rectified frames, respectively;
- **High Framerate**: up to 200 frames per second;
- **Factor Form**: this device is only 70 × 12 × 3 mm and 32g of weight;

Moreover, this sensor is equipped with a fish-eye lens that allows to capture a 150-degree scene from very short distances. The SDK of the Leap Motion device is able to acquire, in addition to infrared images, several hand joints, together with their orientations and bone lengths, as shown in Figure 3.

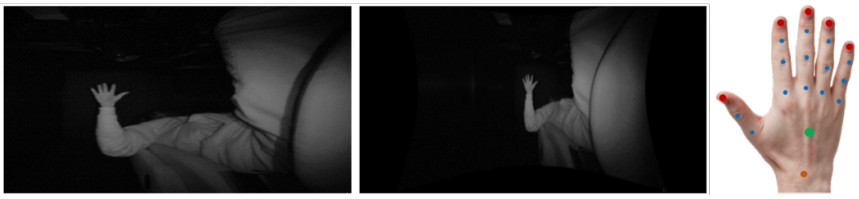

Fig. 3. Data acquired through the *Leap Motion* device: from the left, the raw and the rectified frame, and the hand joints (blue), including the fingertips (red), the palm center (green) and the wrist position (orange).

3.2 Statistics

The *Briareo* dataset contains 12 dynamic gesture classes, designed with a view to the development of an interactive generic system, as follows: *fist* (g00), *pinch* (g01), *flip-over* (g02), *telephone* (g03), *right swipe* (g04), *left swipe* (g05), *top-down swipe* (g06), *bottom-up swipe* (g07), *thumb* (g08), *index* (g09), *clockwise rotation* (g10) and *counterclockwise rotation* (g11).

[4] https://www.leapmotion.com

A total of 40 subjects (33 males and 7 females) have taken part to the data collection. Every subject performs each gesture 3 times, leading to a total of 120 collected sequences. Each sequence lasts at least 40 frames. At the end of this procedure, we record an additional sequence including all hand gestures in a single recording. The three cameras have been synchronized so that the frames at a certain instant depict the same scene.
The following data are released within the dataset: RGB images (traditional camera), depth maps and infrared intensities (Pico Flexx), raw and rectified infrared images (Leap Motion), 3D hand joints (Leap Motion SDK). Samples are reported in Figure 2 and Figure 3.

4 Proposed Baselines

In this Section, we investigate the use of two methods to tackle the gesture classification task, handling the temporal evolution of the dynamic hand gestures in two different ways: 3D convolutions and recurrent neural architecture.

4.1 3D Convolutional Network

Architecture. Taking inspiration from [14], we propose a 3D convolutional neural network to tackle the gesture classification task. Therefore, the temporal evolution of the hand gestures is handled through 3D convolutional layers. We adopt the architecture of C3D [17] which consists of 8 convolutional layers, 5 max-pooling layers and a *softmax* layer. The last 487-dimensional fully connected layer of the original architecture is replaced with a 12-dimensional layer to deal with the number of classes of the *Briareo* dataset.

Training Procedure. Input frames are resized to 112×112 to deal with the C3D architecture constraints and grouped in mini-batches of size 8. As optimizer, we exploit the Stochastic Gradient Descent (SGD) with a learning rate of 10^{-2} and a momentum of 0.5. We use the binary categorical cross-entropy as loss function. Input images are normalized so that the mean and the variance are 0 and 1, respectively. Since each gesture of the dataset has a different duration, we create fixed-length input sequences in the following way: given a single sequence, starting from the central frame (w.r.t. the whole length of the sequence) 20 contiguous frames towards the beginning and 20 towards the end are extracted and stacked to form the input of the proposed architectures.

Table 2. Inference time of the C3D model w.r.t different architectures and input types.

	RGB	**Depth**	**Infrared**
Nvidia 1080 Ti	1.96 ± 0.49 ms	2.26 ± 1.17 ms	2.19 ± 0.93 ms
Nvidia Titan X	1.87 ± 0.77 ms	2.07 ± 0.89 ms	1.86 ± 0.92 ms
CPU	4.01 ± 0.24 s	3.89 ± 0.24 s	3.88 ± 0.24 s

4.2 Long-Short Term Memory

Architecture. Differently from the previous network, here we aim to handle the temporal evolution of the hand gesture through a recurrent neural network, in particular the *Long-Short Term Memory* (LSTM) [7]. The LSTM model employed is described by the following equations:

$$I_t = \sigma(W_i x_t + U_i H_{t-1} + b_i) \tag{1}$$
$$F_t = \sigma(W_f x_t + U_f H_{t-1} + b_f) \tag{2}$$
$$O_t = \sigma(W_o x_t + U_o H_{t-1} + b_o) \tag{3}$$
$$G_t = tanh(W_c x_t + U_c H_{t-1} + b_c) \tag{4}$$
$$C_t = F_t \odot C_{t-1} + I_t \odot G_t \tag{5}$$
$$H_t = O_t \odot tanh(C_t) \tag{6}$$

in which F_t, I_t, Ot are the gates, C_t is the memory cell, G_t is the candidate memory, and H_t is the hidden state. W, U, and b are learned weights and biases, while x_t corresponds to the input at time t as defined in the previous section. Finally, the $\odot$ operator is the element-wise product.
We exploit a LSTM module with a hidden size of 256 units and 2 layers, adding a final fully connected layer with 12 units, corresponding to the number of the gesture classes.

Training Procedure. As reported in Section 3.1, for each frame the Leap Motion SDK gives the 3D joints of the hand and the palm center, represented with the (x, y, z) coordinates in the 3D space. A feature set is then created, including the position of each finger joint and of the palm center, along with the speed and the direction (expressed in terms of *yaw*, *pitch* and *roll* angles) of the fingertips. Data are then normalized to obtain zero mean and unit variance.
The network is then trained using as input this pre-processed data, exploiting the Adam optimizer [8] with learning rate 10^{-3}, weight decay 10^{-4}, and a batch size of 2. As loss function, we use the binary categorical cross-entropy. We empirically set the length of each training sequence equal to 40 frames.

5 Experimental Results

The following experimental results have been obtained in a *cross-subject* setting: we randomly put the recordings of 32 subjects in the train and the validation set and the recordings of the other 8 subjects in the test set. We maintain this division for every test here reported.
Aiming to investigate the contribution of each input modality to the final hand gesture classification accuracy, we train the C3D model separately on the three modalities, *i.e.* RGB, infrared, and depth images. Moreover, we train the proposed LSTM model on the 3D hand joints computed by the Leap Motion SDK. The overall accuracy w.r.t. each gesture and input type is reported in Table 3.
The model that analyzes RGB images obtains the worst result, due to the low

Table 3. Results expressed in terms of accuracy and improvement of the proposed models with respect to the 12 hand gesture classes of the *Briareo* dataset.

Gesture	**Gesture Label**	**RGB**	**C3D Depth**	**Infrared**	**LSTM 3D Joints**
Fist	g0	0.542	0.708	0.750	0.875
Pinch	g1	0.833	0.875	0.958	1.000
Flip-over	g2	0.792	0.750	0.875	0.958
Telephone call	g3	0.625	0.792	1.000	1.000
Right swipe	g4	0.833	0.833	0.917	0.917
Left swipe	g5	0.833	0.917	0.792	1.000
Top-down swipe	g6	0.917	0.750	0.958	1.000
Bottom-up swipe	g7	0.750	0.833	0.875	0.917
Thumb up	g8	0.917	0.625	1.000	0.875
Point	g9	0.667	0.708	1.000	1.000
Rotation (CW)	g10	0.542	0.375	0.750	0.917
Rotation (CCW)	g11	0.417	0.958	0.635	0.875
	all	**0.722**	**0.760**	**0.875**	**0.944**
Improvement		-	+0.038	+0.153	+0.222

brightness and contrast of the acquired images, even though some gestures (*e.g. g6* and *g8*) are easily recognized. As expected, a significant improvement is introduced when analyzing depth maps and infrared images, thanks to the higher image quality and reliability.
As shown in the right part of Table 3, the LSTM model, which analyzed 3D hand joints, achieves the best overall accuracy. However, this performance is based on a correct localization of the 3D hand joints provided by the Leap Motion SDK, limiting the applicability of this method to real world applications.

Considering the high accuracy obtained by the proposed models, we developed a reference framework for the classification of gestures. The C3D model and the infrared images have been selected to deal with both accuracy and speed, without being dependent on external software (*e.g.* the Leap Motion SDK).
The proposed framework processes input data frame by frame and a temporary buffer is maintained, through a sliding windows approach. As soon as 40 frames are stacked, the buffer is classified by the C3D model. The gesture is classified and considered valid only if the prediction confidence reaches a certain threshold, empirically set to 0.85.
In Figure 4 we report the flow chart of the proposed framework, and some screenshots of the graphical user interface, showing infrared and depth frames on the left, and the predicted gesture label on the right.

Finally, we test the inference time of the C3D model in a desktop computer equipped with an *Intel i7-6850K* (3.8 GHz) and 64 GB of memory. This test is carried out on two different GPU, namely the *Nvidia 1080 Ti* and the *Nvidia Titan X*, as well as without graphical accelerators. The model has been developed using *PyTorch*. For investigation purposes, we test the network with each input

type, *i.e.* RGB, infrared, and depth (spatial resolution and data precision vary). The times required for the inference of a single frame are reported in Table 2. As it can be seen, real time performance are achieved when running the model on GPUs.

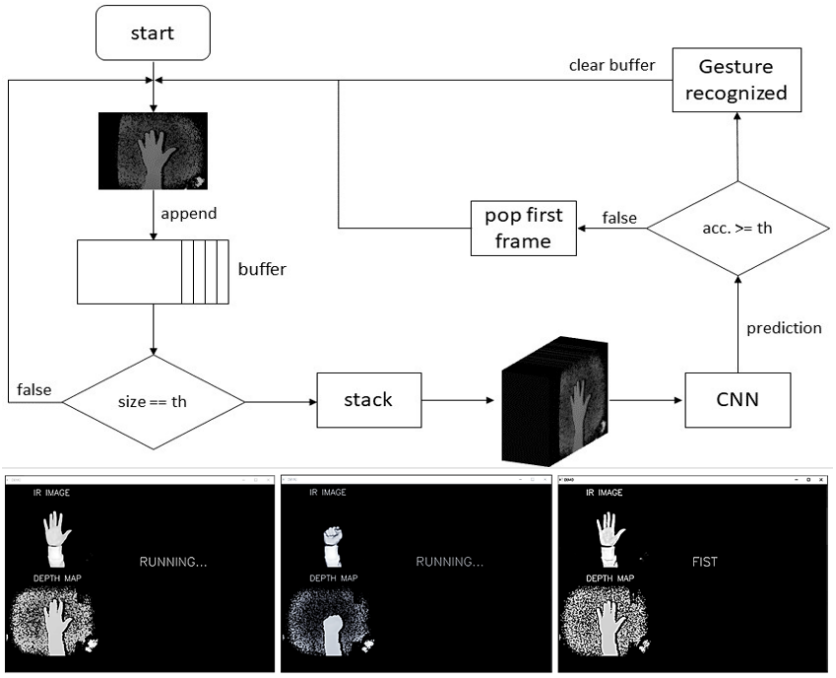

Fig. 4. From the top, the flow chart of the sliding window-based framework for the gesture segmentation and classification task. Then, sample output of the developed framework with the predicted gesture on the right.

6 Conclusions

A new dataset, called *Briareo* and designed for the classification of hand gestures in the automotive setting, has been presented. The dataset contains recording acquired from an innovative point of view with three different acquisition devices, *i.e.* RGB, depth, and infrared cameras.
A C3D-based and a LSTM-based network have been trained and tested on the proposed dataset in order to investigate the quality and the complexity of the collected images and 3D hand joints, achieving a significant accuracy and representing a challenging baseline for future work.
Finally, a real-time hand gesture recognition framework has been presented, showing the capabilities of the proposed dataset and models.

References

1. Borghi, G., Frigieri, E., Vezzani, R., Cucchiara, R.: Hands on the wheel: a dataset for driver hand detection and tracking. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (2018) 1, 3, 4
2. Borghi, G., Gasparini, R., Vezzani, R., Cucchiara, R.: Embedded recurrent network for head pose estimation in car. In: IEEE IV (2017) 1
3. Borghi, G., Vezzani, R., Cucchiara, R.: Fast gesture recognition with multiple stream discrete hmms on 3d skeletons. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 997–1002. IEEE (2016) 2
4. Boulahia, S.Y., Anquetil, E., Multon, F., Kulpa, R.: Dynamic hand gesture recognition based on 3d pattern assembled trajectories. In: IPTA. IEEE (2017) 4
5. Dong, Y., Hu, Z., Uchimura, K., Murayama, N.: Driver inattention monitoring system for intelligent vehicles: A review. IEEE transactions on intelligent transportation systems **12**(2), 596–614 (2011) 1
6. Escalera, S., Baró, X., Gonzalez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: Workshop at the ECCV. pp. 459–473. Springer (2014) 3
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997) 8
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 8
9. Liu, W.: Natural user interface-next mainstream product user interface. In: 2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design 1. vol. 1, pp. 203–205. IEEE (2010) 1
10. Lyznicki, J.M., Doege, T.C., Davis, R.M., Williams, M.A., et al.: Sleepiness, driving, and motor vehicle crashes. Jama **279**(23), 1908–1913 (1998) 2
11. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with leap motion and kinect devices. In: 2014 IEEE ICIP. pp. 1565–1569. IEEE (2014) 3, 4
12. McKnight, A.J., McKnight, A.S.: The effect of cellular phone use upon driver attention. Accident Analysis & Prevention **25**(3), 259–265 (1993) 2
13. Molchanov, P., Gupta, S., Kim, K., Kautz, J.: Hand gesture recognition with 3d convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1–7 (2015) 4
14. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4207–4215 (2016) 3, 4, 7
15. Ohn-Bar, E., Trivedi, M.M.: Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. IEEE transactions on intelligent transportation systems **15**(6), 2368–2377 (2014) 3, 4
16. Sarbolandi, H., Lefloch, D., Kolb, A.: Kinect range sensing: Structured-light versus time-of-flight kinect. Computer vision and image understanding pp. 1–20 (2015) 5
17. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4489–4497 (2015) 7
18. Venturelli, M., Borghi, G., Vezzani, R., Cucchiara, R.: From depth data to head pose estimation: a siamese approach. In: 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2017) (2016) 2