Real-Time Human Body Pose Estimation for In-Car Depth Images

Helena R. Torres, Bruno Oliveira, Jaime Fonseca, Sandro Queirós, João Borges, Nélson Rodrigues, Victor Coelho, Johannes Pallauf, José Brito, José Mendes

Real-Time Human Body Pose Estimation for in-Car Depth Images

Helena R. Torres[1,*], Bruno Oliveira[1,*], Jaime Fonseca[1], Sandro Queirós[1], João Borges[1], Nélson Rodrigues[1], Victor Coelho[3], Johannes Pallauf[2], José Brito[4], José Mendes[1]

[1]Algoritmi Center, University of Minho, Guimarães, Portugal
[2]Bosch, Abstatt, Germany
[3]Bosch, Braga, Portugal
[4]2Ai, Polytechnical Institute of Cávado and Ave, Barcelos, Portugal

Abstract. Over the next years, the number of autonomous vehicles is expected to increase. This new paradigm will change the role of the driver inside the car, and so, for safety purposes, the continuous monitoring of the driver/passengers becomes essential. This monitoring can be achieved by detecting the human body pose inside the car to understand the driver/passenger's activity. In this paper, a method to accurately detect the human body pose on depth images acquired inside a car with a time-of-flight camera is proposed. The method consists in a deep learning strategy where the architecture of the convolutional neural network used is composed by three branches: the first branch is used to estimate the confidence maps for each joint position, the second one to associate different body parts, and the third branch to detect the presence of each joint in the image. The proposed framework was trained and tested in 8820 and 1650 depth images, respectively. The method showed to be accurate, achieving an average distance error between the detected joints and the ground truth of 7.6 pixels and an average accuracy, precision, and recall of 95.6%, 96.0%, and 97.8% respectively. Overall, these results demonstrate the robustness of the method and its potential for in-car body pose monitoring purposes.

Keywords: Autonomous driving, deep-learning, depth images, pose estimation

1 Introduction

In recent years, the concept of autonomous driving vehicles is emerging owing to an increment on the development of advanced driver-assistance systems [1]. In fact, it is expected that fully automated driving will be the next goal for the automobile industry, which will extinguish the role of the driver. Thus, without the necessity of driving the car, the passengers can spend their time doing other types of activities [2]. In this sense, the need for monitoring all occupants in the car becomes crucial to analyze the passengers' behavior and, therefore, to ensure their safety.

There are different types of visual sensors that can be used for monitoring purposes, including monitoring persons through the detection of their pose. The most common sensor consists of RGB cameras that allow to retrieve visual information of the interior

[*]Both authors contributed equally.

of the car and its occupants. However, these sensors only produce 2D images which hampers the interpretation of the human activity that is enrolled in the 3D world and is susceptible to brightness variations [3]. A depth sensor, which is robust to light variations, can be used to solve this problem, as it also provides information about the distance between an object and the camera, giving a 3D information about the scene. However, they not give textural information, which is useful to detect different parts of the body. Recently, several works have focused on the integration of RGB and Depth sensors (RGB-D), which allow merging the advantages of both sensors [4], which could potentially increase the robustness of human pose detection. However, these sensors are more expensive in comparison with RGB and depth sensors.

To detect the human body pose in images acquired with camera sensors and, thus, the activity of the car's occupants, a robust method for human body pose estimation is needed. Specifically for the in-car environment, few strategies were previously proposed. Indeed, the images acquired inside a car have several occlusions, making some joints undetectable. This problem hampers the direct usage of traditional human pose estimation algorithms and makes this task more challenging than open space pose estimation scenes. The focus of the present work was to develop a method for human body pose estimation in depth images acquired inside a car. For it, a state-of-the-art method developed for open space human pose estimation in RGB images, Part Affinity Fields (PAF), was extended for an accurate and robust detection of human body pose inside a car using depth images [5]. The main requirements of the method were its accuracy in detecting the pose along with its robustness to deal with large variability between different people (*i.e.* regarding body shape or size). Moreover, to deal with fast human movements, it was also important to guarantee the real-time capability of the pose detection. Overall, the current work introduces three main contributions: (1) extension of the PAF method for the detection of the presence/absence of a specific joint in the image; (2) a data augmentation strategy for depth datasets; and (3) a dataset for human body pose detection inside a car with depth images.

The rest of the paper is organized as follows. In section 2, the relationship of the present work with industrial and services systems is presented. In section 3, the state-of-the-art for human body pose estimation is detailed. The proposed framework for the human body pose estimation is described in section 4. In section 5, some implementation details related to the proposed method are outlined, being the results presented in section 6. In section 7, the method's performance is discussed, with the main conclusions of this paper presented in section 8.

2 Relationship to Industrial and Service Systems

With the progress in technology, it is expected that autonomous cars will become part of our life, and so, this new paradigm has high influence in industrial and service systems. In fact, partial and fully autonomous driving can bring social-economic advantages [6]. One of the advantages will be an improvement in road safety, since most of current accidents is caused by errors committed by the driver. Moreover, it is also expected that the traffic flow and the mobility could be improved in the autonomous driving scenario. Another advantage of these systems is related to the

increase of the driver's comfort, which by not being focused on driving will have time to do other types of tasks [7]. However, this technological evolution not only concerns personal aspects but also affects public services. In fact, it is expected that shared autonomous vehicles (SAV) will be the next revolution in public transportation. Similarly to personal autonomous vehicles, the SAV also brings several advantages, such as the improvement of passengers' mobility in an easy and economical way [8].

Despite all the above mentioned advantages in autonomous driving, removing the responsibility of the driver/passengers in the driving activity may not be straightforward as it may seem. Indeed, the safety of both car's occupants and vehicle must always be ensured. Regarding personal vehicles, the driver should have the capability to take control of the car when the autonomous driving option is not safe. Concerning the SAVs, the monitoring of the car environment is crucial to maintain the integrity of the vehicle and to ensure the safety of all passengers. This is intrinsically related to the concept of resilient systems, where a system has the capability to recover from perturbations that may affect its normal functioning, allowing the reduction of the vulnerability of the system. A higher capability of maintaining a good performance of the autonomous vehicles can be achieved by monitoring the occupants of the car, being the proposed method of high interest for industrial and services systems.

3 Related Work

To recognize human body pose in images acquired by camera sensors, and, therefore, the activity of the car's occupants, a robust method for human body pose estimation is required. There are two main classes of algorithms for human body pose detection: generative approaches and discriminative approaches. Generative approaches are designed to fit and deform a model to match the image and detect the pose [9]–[12]. Discriminative approaches are designed to learn a mapping from image features to a body pose, using only the information of the image [13]–[17]. Regarding generative ones, the main advantage is the robustness, once these methods fit one or more previous models to the image, allowing to introduce shape prior information to the method. However, this class of methods uses error minimization functions, being susceptible to be trapped in local minima and requiring high computational cost. In opposition, discriminative methods are fast during inference, allowing to more easily achieve real-time detection. Moreover, these methods are capable to deal with large body shape variations. However, they are inherently limited by the amount and quality of the training data. Nevertheless, discriminative approaches, namely deep learning strategies, have shown to be more robust and accurate for human body pose detection than other type of algorithms [18]–[20].

Owing to the success of discriminative approaches for human body pose estimation, several methods have been developed using these type of strategies, specifically deep learning methods. In [21], a graphical model for human pose estimation was used where convolutional neural networks (CNN) were used to learn the different body parts and their spatial relationships (represented by a mixture model over several possible relations). In [22], the graphic model for the human body pose was obtained using the Markov Random Fields approach. Despite its accuracy in human body pose estimation,

the previous methods rely on graphic models which can fail to model complex human poses. To overcome this issue, in [23], a dual source CNN was used to detect both full body and local body parts to achieve the final body pose estimation. In [24], another method for human pose estimation using CNNs was proposed. In this method, a detection followed by a regression cascade strategy was used, where in a first stage heatmaps are inferred to detect the human body parts and then a regression of these heatmaps is performed to learn body relationships. In this work, heatmaps are confidence maps that represent the belief that a particular body part occurs at each pixel of the image. In [5], this heatmaps concept was also used, where a network with two branches was designed to learn both part locations (heatmaps learning) and their associations (Part Affinity Fields - PAF). In this sense, instead of using regression of the heatmaps to learn body relations, the association of the different body parts are simultaneous learned with the heatmap inference.

Although several methods have been proposed for human body pose detection, these methods were applied mostly for pose estimation in open spaces, with the in-car scenario having received little attention in the research world [9], [25], [26].

4 Methods

The main goal of the proposed method was to accurately detect the human body pose of passengers inside a car in depth images. Fig. 1 presents an overview of the proposed method, which is based on the deep learning-based body pose estimation method presented in [5]. The proposed method uses as input a depth image of the driver acquired from a time-of-flight camera (Fig. 1A) and as output the location of each joint of the different human body parts (Fig. 1D). To obtain the joint positions, a CNN is used to simultaneous predict a set of heatmaps (one for each body part joint, Fig. 1B) and a set of part affinity field vectors that represent the association between the different parts (Fig. 1C). Since the in-car environment produces occlusions of some body parts and the Field-of-View (FoV) of commercial depth sensors may not be enough to visualize all joints once the driver stands near the camera, some joints may not be detectable in the images. Thus, the proposed network also predicts if a joint is presented in the image or not (henceforward called as label detection branch), which may boost the method's robustness.

In Fig. 2, the architecture used for the convolutional neural network is presented. As shown in the figure, the first part of the convolutional network consists in the first ten layers of the VGG-19 [27], which are used to perform a first analysis of the image, generating a set of feature maps $\boldsymbol{F}$. Next, the network is split in three branches for the simultaneous learning of the heat maps, the part affinity fields, and the label detection.

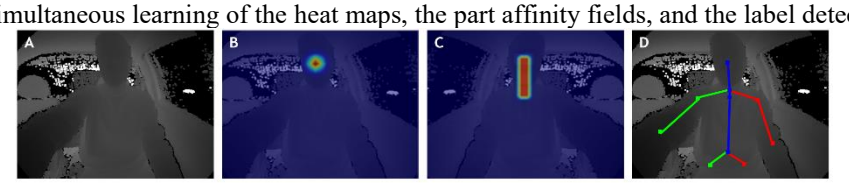

Fig. 1. Overview of the proposed method. (A) Input depth image; (B) Heat map (output of first branch) for the head joint; (B) Part affinity fields (output of second branch) for the association between head and neck joints; (D) final human body pose estimation

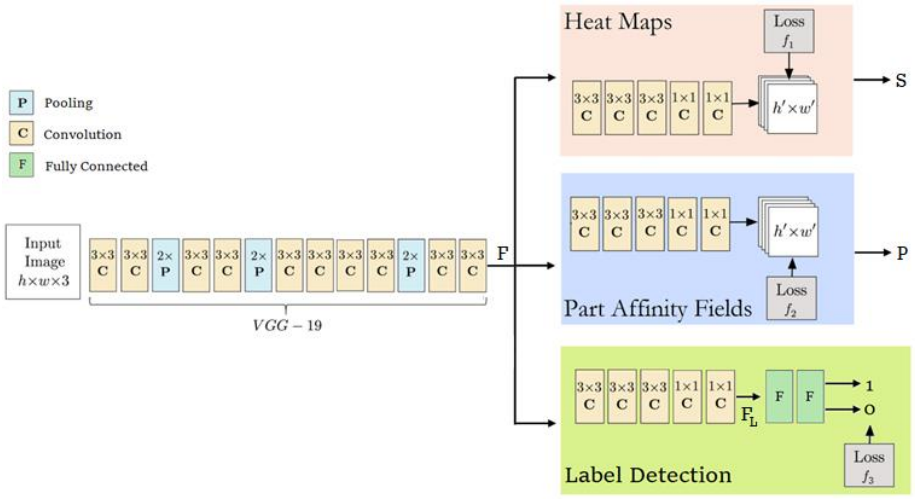

Fig. 2 Architecture of the convolutional network used in the proposed method. The coral branch concerns the learning of the heat maps for body parts' detection and the blue branch concerns the part affinity fields for body part associations. Finally, the green branch is related to the label detection for joint categorization. Adapted from [5].

4.1 Heat Maps for Body Parts' Detection

One of the branches of the convolutional network is used to predict confidence maps of each human body part (coral branch showed in ***Fig. 2***). As previously stated, a confidence map represents the belief that a body joint occurs in a given image pixel, being assigned to each pixel a probability of being a body joint. In this sense, a confidence map can be seen as a 2D gaussian-like function, where the maximum of the gaussian map represents the ideal joint position.

To train the method to predict heat maps, a loss function f_{heat} was applied in the end of this branch to calculate the difference between predictions and ground truth. In this case, the ground truth for the confidence maps was generated using a manual labeling of the joint positions and constructing a gaussian map around the joint locations. The loss function for this branch is given by:

$$f_1 = \sum_{j=1}^{J} ||S_j - S_j^*||, \tag{1}$$

where J represents the number of joints, and S_j and S_j^* are the prediction and ground truth maps for part j, respectively.

In the test phase, the joint position for each body part is given by the maximum of the respective confidence map, *i.e.* its peak, after a non-maximum suppression.

4.2 Part Affinity Fields for Body Part Association

To increase the accuracy of the body part detection, a second branch that measures the association between each pair of body parts was added to the convolutional network (blue branch in Fig. 2). This association is given by part affinity fields, which consists in a vector field between two body joints that encodes the direction between one body joint to another (see Fig. 1). To better understand how the part affinity fields are generated, the reader is kindly directed to [5].

The loss function associated to this branch is given by:

$$f_2 = \sum_{c=1}^{C} ||P_c - P_c^*||, \tag{2}$$

where C represents the number of connections between the different body parts, P_c is the prediction of a part affinity field and P_c^* is the ground truth for the association.

Besides refining the inference of the confidence maps during training, owing to the backpropagation scheme used during it, the part affinity fields are useful when there is more than one person in the image. Indeed, in this scenario, a confidence map by itself may not be enough for an accurate detection because several peaks for the same joint may be detected (one for each person in the image). In this sense, the association between body parts are crucial to understand which joints belong to the same person. However, in this work, and given the FoV of the camera used, we only focused in the detection of one person, and therefore, we only used the part affinity fields to refine the heat map prediction.

4.3 Label Detection for Joint Categorization

Owing to the limited size of an in-car environment and to the reduced FoV of the cameras used for monitoring in this environment, there is a higher probability of certain body joints being outside of the image, specially the extremity limbs (*e.g.* the driver can have its arm outside of the lateral window and the associated joint is therefore not present in the image). It is thus important to understand if the joint is present or not in the image to increase the accuracy of the human body pose estimation. To deal with this problem, we added a third branch to our network. This third branch allow us to categorize the joint with a different label according to its existence in the image, i.e. the joint has the label 0 if it is outside of the image and label 1 if it is inside.

This label detection was achieved by using a set of fully-connected layers to learn non-linear combinations of the features F_L extracted by the convolutional layers (green branch in Fig. 2). Afterwards, a softmax layer was used to assign a probability for the label detection, by taking the output of the fully-connected layers and transforming it into a vector with two prediction scores (one for each class i: presented in the image or not). Each prediction score is given by:

$$p_{i,j} = \frac{e^{a_{i,j}}}{\sum_{k=1}^{K} e^{a_{k,j}}}, i = 1, \ldots, K \tag{3}$$

where a_i represents the output of the last fully-connected layer for class i, K corresponds to the number of classes (in this case $K = 2$), and J represents the number of joints. Please note that the sum of the probabilities is equal to 1 and, therefore, the joint's presence label is given by selecting the class with higher score. The loss function for this branch is given by equation (4), where t_k is the ground truth for the probability of each class (0 or 1).

$$f_3 = \sum_{j=1}^{J} -\frac{1}{K} \sum_{k=1}^{K} \log(p_k, t_k). \tag{4}$$

At test time, the prediction of this third branch is used to verify which heat maps predicted in the first branch should be evaluated. If the label detection branch predicts that the joint is not present in the image, it is considered not detected and the respective heap map is not evaluated. Otherwise, it is assumed that the joint is presented in the image and its position is given by the maximum of the respective heat map, as stated in section 4.1.

The final human body pose estimation is obtained by combining the output of the three branches. Thus, the overall objective function is given by summing the loss of each branch:

$$f = f_1 + f_2 + f_3. \tag{5}$$

5 Experiments

5.1 Dataset Creation

Due to the inexistence of public datasets of depth images in an in-car scenario, it was needed to create our own dataset. Owing to the deep learning nature of the method, a massive amount of data is needed as training data, and therefore, it was needed to create a large dataset. Moreover, besides the high number of training images, the training dataset must also be variable enough to include the large number of actions possible in this scenario. In fact, the accuracy of the method is very dependent of the quality of the dataset. In this sense, we constructed our dataset by acquiring depth images using a time-of-flight camera placed near the windshield in front of the driver. For the construction of the dataset, ten different cars were used to achieve the desired variability in terms of image background. For each one of the ten cars, five subjects acted as driver, performing different actions inside a car (*e.g.* driving, putting the seat belt, picking up the phone, and others) to give the robustness needed for the dataset. The combination of the different cars and different drivers allowed to construct a dataset with 12200 depth images. The dataset was then divided in training, validation, and testing set with 8820, 1730, and 1650 images each, respectively. In this work, the training dataset was used to train the method, the validation set was used to test the progress of the performance of the method during training, and the testing set was used for the final

validation of the proposed method. Note that the cars in each set of images differ from each other to achieve an unbiased evaluation.

Concerning the ground truth for the different body parts, it was constructed by manual labelling of the joint positions. Fourteen joints were used, namely: head (H), neck (N), right/left shoulder (RS/LS), right/left elbow (RE/LE), right/left wrist (RW/LW), chest (C), pelvis (P), and right/left hip (RH/LH). Note that the performance of the method was only evaluated for these upper body joints once the lower body parts are naturally occluded when simulating a driving position. Moreover, the joint categorization (present or not) was also manually defined per image.

5.2 Data Augmentation

As above-mentioned, the ground-truth for the training dataset was obtained manually, which could represent a limitation in respect to the number of training images once manual labelling is a tedious and time-consuming task. In this sense, besides the real dataset constructed, and as common in deep learning strategies, a data augmentation layer was implemented, allowing to generate more training images that the ones initially labelled. Moreover, this data augmentation strategy allowed to increase the variability of the training dataset. Traditionally, data augmentation strategies rely on image flip, rotation, and scaling. Although it is an effective way of increasing image variability during training, such strategies do not modify the intensity information of the image. Although such feature is not so problematic for RGB images, it can for depth ones, as changing the intensity of these images can be a useful way to simulate different camera positions in the world (i.e. the distance between the camera and objects). In this work, the implemented data augmentation layer can simulate these changes in terms of camera's position, allowing to create images where the objects (i.e., the driver) are closer or farther from the camera than they were in fact.

The first step of our data augmentation approach is to convert the 2D depth image in a 3D point-cloud, using the extrinsic parameters of the camera. Upon obtaining the 3D camera coordinates, all points can be transformed by applying a given translation, moving the 3D point cloud in any direction and in any axis. The final step consists in using the intrinsic camera parameters to transform the translated point-cloud into a new 2D depth image. In Fig. 3, it is possible to visualize two examples of the result of our data augmentation strategy: one simulates the camera closer to the driver (Fig. 3B) and the other simulates the camera located farther from the driver (Fig. 3C).

Fig. 3 Data augmentation strategy. (A) Original depth image; (B) Augmented image that simulates the positioning of the camera closer to the driver; (C) Augmented image that simulates the camera being located farther from the driver

5.3 Implementation Details

In the implementation of any deep learning method, the definition and optimization of certain training parameters for each specific problem are fundamental and can have a significant importance in the final accuracy of the method. In this work, experiments were carried out to get the optimal settings for these parameters. Concerning the learning rate and the momentum, these parameters were experimentally set to 0.0004 and 0.9, respectively. For the model optimization, the Adam solver was used with a regularization term of 0.01. For the batch size, 10 images were used. Note that these parameters were chosen by evaluating the method in the validation dataset.

6 Results

One important task to correctly implement a deep learning strategy is to evaluate the progress of the training in the validation dataset. Besides being useful to conduct experiments related with the best parameters to be used in the deep learning strategy, the validation dataset is also needed to avoid problems like overfitting to the training data, which may cause failure of the method when applied in a different set of images. In this sense, evaluating the method's performance during training in the validation dataset allow us to detect when the training converges, avoiding overfitting problems. Figure 4 presents an example graph showing the progression of the loss during training in both training and validation datasets.

To evaluate the performance of the proposed method in the testing set, different evaluation metrics were used. The first metric consists in the distance (D) between the detected joint and the ground truth in pixels. Moreover, the method's accuracy (Ac, percentage of correctly detected joints), the precision (Pr, fraction of correctly detected joints among all the correctly or wrongly detected joints) and the recall (Re, fraction of correctly detected joints among all the correctly detected joints and the wrongly non-detected joints) were also evaluated, being defined by equations (6) to (8).

$$Ac = \frac{T_P + T_N}{N} \times 100 \tag{6}$$

$$Pr = \frac{T_P}{T_P + F_P} \times 100 \tag{7}$$

$$Re = \frac{T_P}{T_P + F_N} \times 100 \tag{8}$$

where T_P, T_N, F_P, and F_N correspond to the true positives, true negatives, false positives, and false negatives, respectively, and N is the total number of testing images.

Table 1 summarizes the method's performance for each joint on the testing dataset. The presented values for the distance metric correspond to the median of the distance errors obtained. Note that the information if a joint is correctly detected or not used in the estimation of the accuracy, precision, and recall is given by the third branch of the proposed network. Moreover, the results showed in the table were obtained using the

model obtained in the ideal epoch for early training stopping, which were estimated using Fig. 4. In Fig. 5, the boxplots for the distance metric are presented to analyze the detection performance distribution in terms of percentiles and outlier points. Fig. 6 presents the evolution of the accuracy, precision, and recall for different distance thresholds (i.e. not using the information provided by the label categorization). In this case, a joint was considered correctly detected only if its distance error is bellow a given threshold. Finally, in Fig. 7, some example results of the proposed strategy for human body pose estimation are presented.

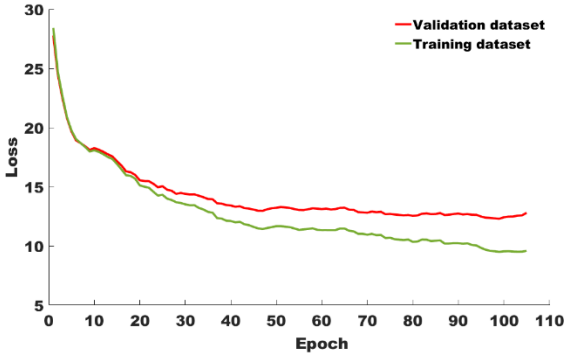

Fig. 4- Method's performance (loss in function of epochs) during training in the training (green line) and validation (red line) datasets.

Table 1- Method's performance in the testing dataset, assessed in terms of distance error (D, pixels), accuracy (Ac, %), precision (Pr, %), and recall (Re, %).

	H	N	RS	RE	RW	RS	RE	RW	C	P	RH	LH	Mean
D	6,0	5,1	5,4	6,7	10,8	6,7	6,4	10,0	9,2	7,2	8,6	9,1	7,6
Ac	98,8	99,6	99,8	97,8	85,0	98,5	96,8	88,1	99,1	98,8	92,3	92,4	95,6
Pr	99,1	99,7	99,8	98,7	82,6	98,8	97,8	83,9	99,4	99,0	96,8	96,6	96,0
Re	99,7	99,9	100,0	99,1	96,0	99,8	98,9	90,4	99,8	99,8	95,1	95,5	97,8

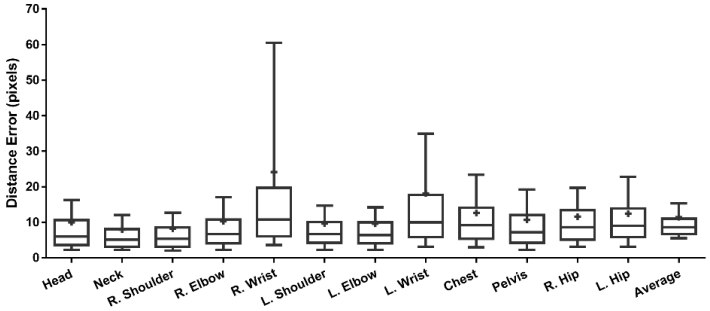

Fig. 5- Boxplots for the distance errors obtained by the proposed method per joint. The ends of the whiskers represent the 10th and 90th percentiles and the crosses represent the mean values.

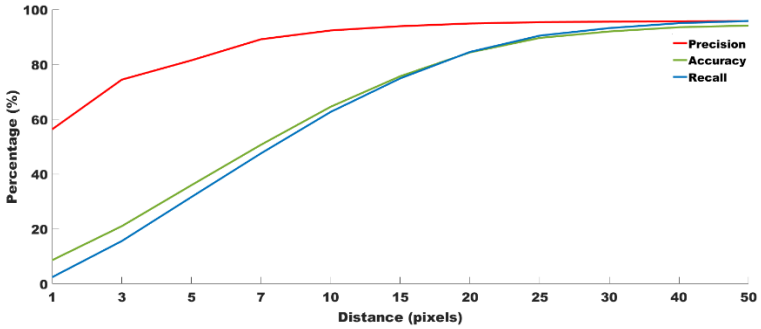

Fig. 6 Variation of accuracy (green line), precision (red line), and recall (blue line) for different distance thresholds in pixels.

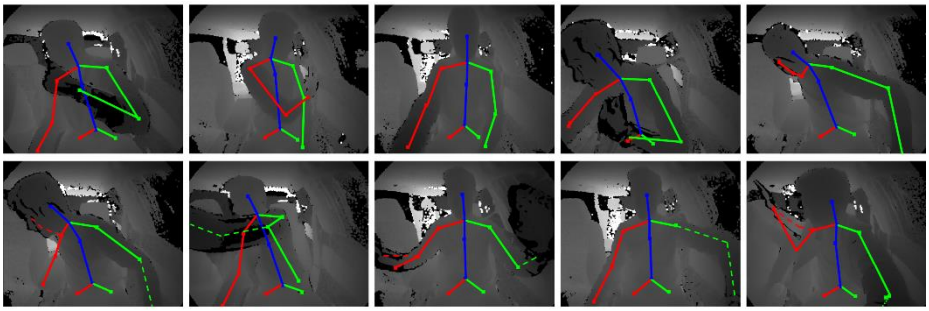

Fig. 7 Qualitative results of the proposed human body detection method. The first row presents examples of good results. In the second row, some examples where the pose estimation failed for a few joints are illustrated, with the ground truth pose shown in dashed lines

7 Discussion

This paper proposed a method for human body pose estimation in an in-car scenario. As stated, an important study to be performed during the implementation of a deep learning strategy concerns the evolution of the training. For that, its performance during training must be evaluated both in the training and validation datasets. Analyzing Fig. 4, it is possible to visualize that the ideal timing for stopping the training would be approximately around the 60th epoch. After this point of the training, the graph suggests that the model may start suffering from overfit to the training dataset, resulting in a very good performance for the images presented in this dataset but a lower or equal performance in the validation dataset.

After analyzing the ideal epoch for early stopping, the model obtained in this epoch was used for the final validation of the method in the testing set. Analyzing Table 1 it is possible to verify the good performance of the human body pose detection method, with a mean distance error for all joints of 7.6 pixels. Moreover, high values for the accuracy, precision, and recall were obtained, showing the method's robustness. The

worst results were obtained for the wrist joints. This can be explained by the fact that these joints are frequently near the image's limits, which can lead to a lower accuracy of the deep learning strategy. Moreover, owing to the proximity of the camera to the driver, these joints are not always present in the camera' field-of-view, which hampers the training process for these joints. This less accurate detection for these extremity joints can also be visualized in Fig. 5, where it is possible to verify the larger mean and interquartile range for these joints. Nevertheless, Fig. 5 corroborates the overall good performance of the method, with narrow boxplots being obtained for the majority of the joints. In Fig. 6, it is shown that the accuracy, precision, and recall are improved with the increase of the threshold for the distance, as expected. Analyzing this graph, one can verify that for a threshold distance of 15 pixels, these evaluation metrics present acceptable values.

Concerning the computational time required by the method, the proposed human body pose detection methods takes approximately 80 milliseconds per image, which gives a performance of 12 frames per second, which proves the nearly real-time capability of the method. Note that this runtime analysis was performed using a laptop with a NVIDIA GeForce GTX-1060 GPU.

One important aspect to take into account in the proposed method is its application in a real setup. In fact, to achieve a real-time estimation of the human body pose inside a car, a high computational power is needed, which can be a limitation for an in-car scenario. However, the constant growth in computation capability of the technology will allow the application of the proposed method in a real scenario. Another aspect to take into account is that the application of software in autonomous cars should follow some existing standards to ensure its applicability in a real scenario. Nevertheless, all these issues will be address in the future.

8 Conclusions and Future Work

This paper presented a framework to detect the human body pose in depth images acquired inside a car. This method consists in a deep learning strategy where a new convolutional network configuration with three branches was used for simultaneous learning of confidence maps for each joint position, body parts' associations, and joint categorization (regarding its existence in the image). The proposed framework was validated in 1650 depth images, achieving an average distance error of 7.6 pixels and an average accuracy, precision, and recall of 95.6%, 96.0%, and 97.8% respectively. Overall, the proposed human body pose estimation method proved to be successful and accurate to detect the driver's pose, while showing a nearly real-time performance.

In future work, we intend to expand the human body pose estimation to more than one person, allowing to monitor not only the driver but all the car' occupants. In addition, we also intend to modify the method to exploit the 3D information provided by the depth image to infer the 3D joint position.

References

[1] J. Levinson *et al.*, "Towards Fully Autonomous Driving: Systems and Algorithms," in *IEEE Intelligent Vehicles Symposium*, 2011, pp. 163–168.

[2] V. A. Banks and N. A. Stanton, "Analysis of driver roles : Modelling the changing role of the driver in automated driving systems using EAST Analysis of driver roles : modelling the changing role of the driver in automated driving systems using EAST," *Theor. Issues Ergon. Sci.*, pp. 1–17, 2017.

[3] D. Regazzoni, G. De Vecchi, and C. Rizzi, "RGB cams vs RGB-D sensors : Low cost motion capture technologies performances and limitations," *J. Manuf. Syst.*, vol. 33, no. 4, pp. 719–728, 2014.

[4] L. Shao, J. Han, D. Xu, and J. Shotton, "Computer Vision for RGB-D Sensors : Kinect and Its Applications," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1314–1317, 2013.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[6] B. Y. S. M. Casner, E. L. Hutchins, D. O. N. Norman, and A. C. Promise, "The Challenges of Partially Automated Driving," in *COMMUNICATIONS OF THE ACM*, 2016, pp. 70–77.

[7] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles : opportunities , barriers and policy recommendations," *Transp. Res. PART A*, vol. 77, pp. 167–181, 2015.

[8] R. Krueger, T. H. Rashidi, and J. M. Rose, "Preferences for shared autonomous vehicles," *Transp. Res. Part C Emerg. Technol.*, vol. 69, pp. 343–355, 2016.

[9] D. Demirdjian and C. Varri, "Driver pose estimation with 3D Time-of-Flight sensor," in *2009 IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems*, 2009, pp. 16–22.

[10] M. Ye and R. Yang, "Real-time Simultaneous Pose and Shape Estimation for Articulated Objects Using a Single Depth Camera," in *CVPR*, 2014, pp. 2345–2352.

[11] M. Ye, Xianwang Wang, R. Yang, Liu Ren, and M. Pollefeys, "Accurate 3D pose estimation from a single depth image," in *2011 International Conference on Computer Vision*, 2011, pp. 731–738.

[12] M. Sigalas, M. Pateraki, and P. Trahanias, "Full-Body Pose Tracking?The Top View Reprojection Approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1569–1582, Aug. 2016.

[13] J. Shotton *et al.*, "Real-Time Human Pose Recognition in Parts from Single Depth Images," *Commun. acm*, vol. 56, no. 1, 2013.

[14] J. Shotton *et al.*, "Efficient Human Pose Estimation from Single Depth Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

[15] M.-H. Tsai, K.-H. Chen, and I.-C. Lin, "Real-time upper body pose estimation from depth images," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 2234–2238.

[16] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter, and C. Pantofaru, "An adaptable system for RGB-D based human body detection and pose estimation," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 39–52, Jan. 2014.

[17] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, "Towards Viewpoint Invariant 3D Human Pose Estimation," Springer, Cham, 2016, pp. 160–177.

[18] V. Belagiannis, A. Zisserman, and V. G. Group, "Recurrent Human Pose Estimation," in *12th IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.

[19] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-Context Attention for Human Pose Estimation," in *arXiv preprint arXiv:1702.07432*, pp. 1831–1840.

[20] M. R-cnn, P. Doll, and R. Girshick, “Mask R-CNN,” in *Computer Vision (ICCV)*, 2017, pp. 2980–2988.
[21] X. Chen and A. Yuille, “Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations,” in *Conference on Neural Information Processing Systems*, 2014, pp. 1–9.
[22] J. Tompson, A. Jain, Y. Lecun, and C. Bregler, “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1–9.
[23] X. Fan, K. Zheng, Y. Lin, and S. Wang, “Combining Local Appearance and Holistic View : Dual-Source Deep Neural Networks for Human Pose Estimation.”
[24] A. Bulat and G. Tzimiropoulos, “Human pose estimation via Convolutional Part Heatmap Regression,” in *European Conference on Computer Vision*, 2016.
[25] G. Borghi, “POSEidon : Face-from-Depth for Driver Pose Estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5494–5503.
[26] P. Murthy, O. Kovalenko, A. Elhayek, C. Gava, and D. Stricker, “3D Human Pose Tracking inside Car using Single RGB Spherical Camera,” 2017.
[27] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *Computer Vision and Pattern Recognition*, 2014, pp. 1–14.