Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 204 (2022) 479–487

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry Sciences and Computer Science Innovation

Optimized in-vehicle multi person human body pose detection

Sandra Dixe[a,*], João Sousa[b], Jaime C. Fonseca[a], António H. J. Moreira[c], João Borges[d]

[a]*Algoritmi Center, University of Minho, 4800-058 Guimarães, Portugal*
[b]*Engineering School, University of Minho, 4800-058 Guimarães, Portugal*
[c]*2Ai – School of Technology, IPCA, 4750-810, Barcelos, Portugal*
[d]*Polytechnic Institute of Cávado and Ave, 4750-810 Barcelos, Portugal*

Abstract

The number of Shared Autonomous Vehicles (SAV) will increase in the coming years. The absence of human driver will create a new paradigm for in-car safety. This paper addresses this problem by presenting an approach to estimate the human body pose inside a vehicle. We propose to use a customized version of the OpenPose framework, to perform the task of human body pose detection for the front passengers inside a vehicle. The OpenPose method was been evaluated with three different backbones: VGG19, MobileNetV1 and MobileNetV2, using different hyperparameters and ablation scenarios. Moreover, synthetic images were used, which simulate a depth sensor perspective from the center of the front seats. The dataset is comprised by images with 1 and 2 passengers, from 18 different subjects inside of 7 different vehicles, thus making a total of 45360 different images. The OpenPose method with the MobileNetV2 backbone showed the most efficient results between precision and performance, achieving a mean Average Precision (mAP) of 90%, Area Under ROC Curve (AUC) of 73%, and 0.0189 seconds per image (s/img).

Keywords: Shared Autonomous Vehicles; Deep Learning; Computer Vision; Body Posture Detection.

1. Introduction

Shared autonomous vehicles (SAVs) present themselves as the transport approach of the future, with a positive relationship between cost and safety. But the fact that there is no driver, as someone responsible for the environment, can create safety issues. The quality of service can be directly or indirectly hampered by passengers, for example, their behavior is unpredictable and can generate many risks. Hence the need arises to study the human pose of passengers inside the car in order to monitor their behavior and subsequently be able to detect human actions and predict them optimally. To ensure the safety of passengers and the monitoring of the interior of a SAV, several works have been developed. Torres *et al.* [1] proposed a passenger monitoring system using a Deep Learning (DL) strategy to accurately

* Corresponding author.
E-mail address: b12147@algoritmi.uminho.pt

1877-0509
10.1016/j.procs.2022.08.059

detect the driver human body pose. DL strategies require a considerable amount of data, Borges *et al.* proposed tools for automatically generating synthetic [2] and real [3] in-car datasets for human body pose detection. The synthetic dataset presents a custom vehicle environment that simulates humans, sensors, and car models, however, these may lack some of the realism from a real dataset. Furthermore, the real dataset approach combines optical and inertial-based systems to achieve in-car motion capture.
In this paper, a solution based on DL techniques was developed for body pose detection in order to overcome this existing problem in SAVs, thus providing relevant features for action recognition algorithms, which focus on the detection of violent actions among passengers [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Moreover, a dataset of synthetic images, based on depth frames, inside vehicles was used. The OpenPose method was evaluation through different permutations of backbones, hyperparameters and ablation configurations. Finally a specific OpenPose configuration was identified. The main contributions of the paper are as follows:

- OpenPose precision and performance study in regard to multiple backbone, hyperparameter, ablation permutations;
- OpenPose architecture for multi passenger in-vehicle body pose detection.

The rest of the paper is organized as follows. Section 2 presents the state-of-the-art of different methods for human body pose estimation in different environments. The dataset creation and model implementation is described in section 3. Experiments are described in section 4, with their corresponding results. The discussion is presented in section 5. In section 6, the paper is concluded and some future work presented.

2. Related Work

Several studies have focused on the detection of human body pose in various environments, and for one or multiple persons. Toshev *et al.* [4] presented in 2014 one of the first major methods to apply deep learning techniques to estimate body pose. The prediction of body pose is formulated as a linear regression problem and outperformed all existing models so far. Demirdjian *et al.* [20] presented in 2009 the problem of vision-based driver pose estimation, i.e., estimating the location and orientation of the driver's limbs. For pose estimation, a variant of the articulated ICP algorithm, a 3D model fitting approach, was proposed, which is able to incorporate the uncertainty in visual observation and model data into the pose estimation framework. Ye *et al.* [22] presents a new system for estimating body pose configuration from a single depth map. It combines pose detection and pose refinement. Ye *et al.* [21] presents in 2014 a new real-time algorithm for simultaneous pose and shape estimation for articulated objects, such as humans and animals. The novelty consisted in the pose estimation component incorporating the articulated deformation model with parameterization based on exponential maps into a Gaussian mixture model. Sigalas *et al.* [23] presents a model-based approach for extracting and tracking markerless articulated whole-body poses and tracking in RGB-D sequences. The performance of the proposed method was compared with Microsoft's Kinect SDK and NiTE, and the results validated the approach presented by the author. Shotton *et al.* presents two papers for the task of body pose estimation [24, 25]. [24] proposed a method to quickly and accurately predict 3D positions of body joints from a single depth image without using temporal information. The evaluation shows high accuracy on synthetic and real test sets and investigates the effect of various training parameters. [25] presents two new approaches for human pose estimation. Both can quickly and accurately predict the 3D positions of body joints from a single depth image without using any temporal information. The key to both approaches is the use of a large, realistic and highly varied synthetic set of training images. Tsai *et al.* [26] proposes the use of a new two-stage method to estimate the probability maps of users' upper body parts. These maps are then used to evaluate the fitness of the body part region for pose retrieval. It was shown to have satisfactory real-time results with a moderate size of training data. Buys *et al.* [27] presents a new method for generating training data of human postures with attached objects. The results showed a significant increase in body part classification accuracy for subjects from 60% to 94% using the generated image set. Su *et al.* [28] proposed a cascade type feature aggregation network which consists of several hourglass networks for obtaining more accurate human body pose estimation in terms of robustness to partial occlusion and low resolution. Their proposed network works with a ResNet backbone and executes feature aggregation by using features attained in previous stages as an input for current one, fusing all estimated body heatmaps derived from all stages to enhance network

performance. They evaluated their model on MPII and LIP datasets, obtaining an accuracy of 93.9% and 91% for total joints, respectively. Cao *et al.* [17] presented a multi-stage Convolutional Neural Network (CNN) based architecture for multi-person 2D body pose detection. The proposed architecture consisted of two branches: the first branch, estimates the confidence map of body joints, and the second one predicts the Part Affinity Field (PAF) that represent the links between joints, and the final prediction will be made by fusing the outputs of the two stages. Evaluations done on COCO 2016 and MPII datasets shows good accuracies; 68.2% and 79.7%, respectively. Although several methods have been proposed for human body pose detection, these methods have mainly been applied for pose estimation in open spaces, with the in-car scenario receiving little attention in the research world [28, 29, 30].

3. Implementation

This work aims to study the detection of human body pose for front seat passengers inside vehicles, to be subsequently used in monitoring systems for detecting passenger actions inside SAVs. Thus our focus is to reach a better understanding between algorithmic configurations, precision and performance. For this purpose, all implementations were based on the OpenPose method, which was trained and tested on a synthetic dataset with depth images of one and two people inside several vehicles. The tool used for the generation of our dataset was presented by Borges *et al.* [2].

3.1. Dataset Generation

The generated dataset, from the toolchain [2], is comprised by NST Depth frame, $b_{x,y}$, features (as shown in Figure 1), and 20 human body joints per-pixel position, i.e. labels (Figure 2). The synthetic dataset includes data generated for one (Figure 1 (a)) and two passengers (Figure 1 (b)), using 7 car models, N_{cm}, and 18 subjects, Z, with associated Gaussian poses, N_{hgp}, for a total of 22680 samples for one passenger and a total of 22680 for two passengers. Care was taken to ensure that no identical subjects were ever crossed in the generation of the two-subject dataset, thus ensuring variability.

3.2. Model OpenPose

From the state-of-the-art models, the OpenPose model was adopted [17]. There were several factors that contributed to the choice of this method, one of the main reasons being the fact that it is one of the most successful methods for estimating human body pose and, moreover, being open source [18]. This method presents very good results in estimating the pose of multiple subjects in real time, this robustness is relevant in this scenario because the space where the subjects are is reduced, i.e. inside the vehicle, and this generates added difficulties in the detection of their body pose. The architecture of OpenPose consists of an input auxiliary CNN (i.e. backbone) that extracts a

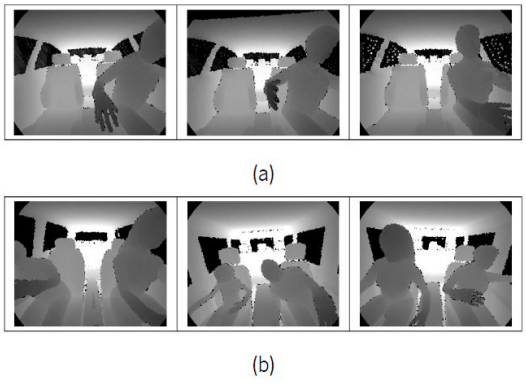

Fig. 1. Representation of NST Depth frames (resolution is 352x287 pixels, with a depth range of 2.5m) that constitute the dataset. a) Examples of images of a person inside a vehicle; b) Examples of images of two people inside a vehicle.

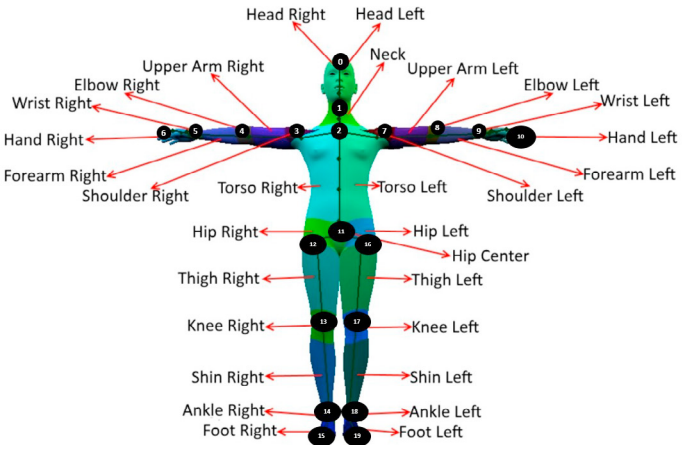

Fig. 2. Default human model labels presented in [2]. A total of 20 body part segmentation labels are used and enumerate by the author, and are directly related with human body joints and body segments that represent joint connectivity. In our work, 20 joints are defined and used as shows in each circled number.

set of feature maps from the input data, these feature maps, F, are used as input data for the two-branch network: one to estimate the confidence maps, *heatmaps*, of the positions of the body joints, where the set $S = (S1, S2, ..., SJ)$ has J confidence maps, one per human body joint, where $S_j \in R^{w \times h}$, $j \in \{1...J\}$; and another that estimates the PAF, i.e. the association/linkage between the different joints, where the set $L = (L1, L2, ..., LC)$ has C vector fields, one per limb, where $Lc \in R^{w \times h \times 2}$, $C \in \{1...C\}$ (shown in Figure 2). An example of a label for a confidence map is for instance joint 1, $S1$, in Figure 2, an example of a vector field, $C1$, is for instance the link $S1 - S2$ in Figure 2. In this work we consider a total of 20 confidence maps, $J = 20$, represented in Figure 2.

In this work we propose to study the application of this method with different backbone configurations, thus aiding to improve precision and performance metrics for in-vehicle human body pose detection. Figure 3, on the right side, shows the arquitecture of the OpenPose model and respectively the backbones used in this study. Backbone configurations are as follow: CMU consists of the first 10 layers of VGG19; MobileNetV1 consists of the first 8 layers with added concatenated features; and MobileNetV2 consists of the first 14 layers with added concatenated features. The 3 backbones used for the study of the OpenPose method are highlighted in green in Figure 3.

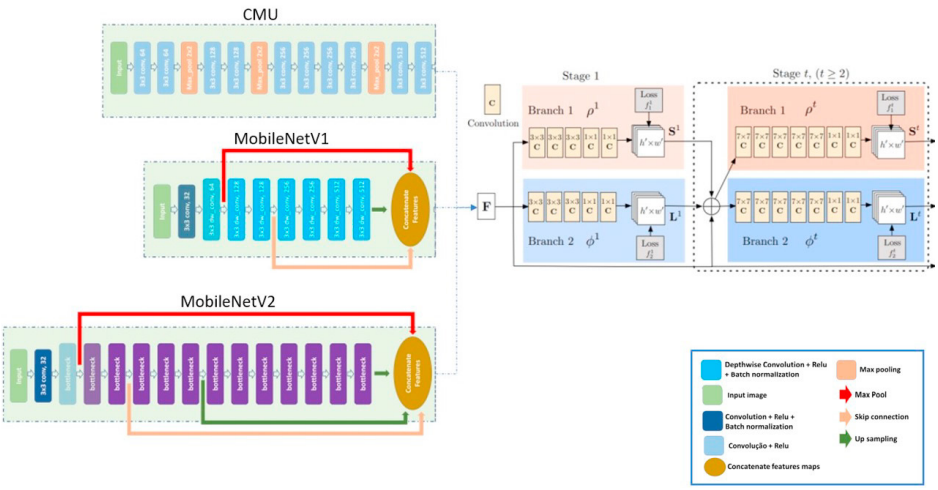

Fig. 3. Schematic of the OpenPose model and respective backbones used in this study.

4. Experiments

All tests were performed in Tensorflow 1.4.1, using the OpenPose [18] source code, running on an Intel (R) Xeon (R) Gold 6140 CPU 2.30 Ghz processor, with 128GB RAM and 1x NVIDIA Tesla V100-PCIE-16GB GPU.

4.1. Synthetic Dataset

The synthetic dataset generated, i.e. using the specifications mentioned in Section 3.1, was splited in train, validation and test parts, as shown in Table 1. Thus avoiding the presence of the same subjects between train, validation and test datasets.

Table 1. Dataset division for evaluations.

	Train	**Validation**	**Test**
Number of Subjects	12	4	2
Subjects Id	$Z_{1:6;13:18}$	$Z_{7:8;11:12}$	$Z_{9,10}$
Number of Images	33684	11228	5614

4.2. Evaluations

To better understand which is the best model configuration (i.e. ablation and hyperparameter), in terms of precision and performance, an iterative evaluation procedure was defined, due to its training time consuming nature. Moreover, a first extensive evaluation was done with the default backbone configuration, the CMU, $EV1$. Based on the best combination of parameters obtained for the previous backbone, it was then defined an additional set of evaluations with the MobileNetV1 and MobileNetV2 backbones, $EV2$ and $EV3$ respectively. All evaluations used a maximum number of epochs equal to 30, batch size of 16, an initial learning rate of 0.001 and a learning rate decay of 0.0005.

4.2.1. EV1: CMU

In order to extend the evaluations to the different ablation and hyperparameter parameters, with a reduced evaluation time penalty, isolated parameter evaluations were defined while considering a base permutation, $EV1_1$ (as shown in Table 2). Moreover, *Aug* represents the enabling of spatial transformation techniques (i.e. translation, rotation, zoom), omiting traditional RGB augmentation due to the input images being depth based. *Image* represents the input image resolution, σ controls the spread of the peak of the gaussian heatmap of each joint ground-truth position, R-LR-p represents the enabling of Reduce-LearningRate-plateau technique [19], and T represents the number of stages of refinement, $t \in \{1, ..., T\}$ (as shown in Figure 3). All evaluations results are presented in Table 4 and Figure 4.

4.2.2. EV2: MobileNetV1 and EV3: MobileNetV2

Following $EV1$ evaluations, the best configuration was selected, $EV1_7$. Morevover, its parameter configurations were fixed for all $EV2$ and $EV3$ evaluations. For $EV2$ and $EV3$, MobileNetV1 and MobileNetV2 respectively, other hyperparameters were added to help manipulate and enhance the evaluations metrics (as shown in Table 3). These parameters are $Conv1$, $Conv2$ and the feature map, F, output size, F_{xy}.

Where $Conv1$ defines a percentage of the original number of filters in the convolutional layers of the backbone, which is used to extract the feature maps, F, $Conv2$ defines a percentage of number of filters in the convolutional layers of each branch from the OpenPose architecture network (i.e. for all stages, t), and the parameter F_{xy} defines the expected size of the backbone output, F, that will serve as input to the two branches of the OpenPose architecture.

All evaluations results are presented in Table 4 and Figure 4.

Table 2. EV1: CMU evaluations. Where *Augmentation* represents the enabling of spatial transformation techniques, omiting traditional RGB augmentation due to the input images being depth based. *Image* represents the input image resolution, σ controls the spread of the peak of the gaussian heatmap of each joint ground-truth position, R-LR-p represents the enabling of Reduce-LearningRate-plateau technique [19], and T represents the number of stages of refinement, $t \in \{1, ..., T\}$ (as shown in Figure 3). Parameters in bold represent the different permutations from the base evaluation $EV1_1$.

	Aug	**Image**	σ	**R-LR-p**	**T**
$EV1_1$	**FALSE**	**368x304**	**8**	**FALSE**	**1**
$EV1_2$	**TRUE**	368x304	8	FALSE	1
$EV1_3$	FALSE	368x304	8	**TRUE**	1
$EV1_4$	FALSE	368x304	8	FALSE	**2**
$EV1_5$	FALSE	368x304	8	FALSE	**6**
$EV1_6$	FALSE	**304x256**	8	FALSE	1
$EV1_7$	FALSE	**256x208**	8	FALSE	1
$EV1_8$	FALSE	368x304	**4**	FALSE	1
$EV1_9$	FALSE	368x304	**6**	FALSE	1
$EV1_{10}$	FALSE	368x304	**10**	FALSE	1
$EV1_{11}$	FALSE	368x304	**12**	FALSE	1
$EV1_{12}$	FALSE	368x304	**14**	FALSE	1

Table 3. *EV*2: MobileNetV1 and *EV*3: MobileNetV3 evaluations. Where *Conv*1 defines a percentage of the original number of filters in the convolutional layers of the backbone, which is used to extract the feature maps, F, and *Conv*2 defines a percentage of number of filters in the convolutional layers of each branch from the OpenPose architecture network (i.e. for all stages, t), and the parameter F_{xy} defines the expected size of the backbone output, F, that will serve as input to the two branches of the OpenPose architecture, and T represents the number of stages of refinement, $t \in \{1, ..., T\}$ (as shown in Figure 3). Parameters in bold represent the different permutations from the base evaluation $EV2{:}3_1$.

	*Conv*1	*Conv*2	F_{xy}	**T**
$EV2{:}3_1$	**1**	**1**	**28x28**	**1**
$EV2{:}3_2$	1	1	28x28	**2**
$EV2{:}3_3$	1	1	28x28	**6**
$EV2{:}3_4$	**1.4**	1	28x28	1
$EV2{:}3_5$	**1.4**	**1.4**	28x28	1
$EV2{:}3_6$	**0.75**	**0.75**	28x28	1
$EV2{:}3_7$	**0.75**	**1**	28x28	1
$EV2{:}3_8$	1	1	**56x56**	1

5. Discussion

This paper proposes the use of the OpenPose model to estimate the human pose inside of vehicles. Different backbone, ablation and hyperparameter experiments were performed, in order to obtain the best compromise between precision and performance. Based on the results obtained in the different evaluations (Table 4 and Figure 4), it is possible to conclude that the trained models perform well in estimating the location of the joints of the subjects inside the vehicles. Besides the sensor position, the vehicle structure itself hinders the visibility of some joints, such as the joints in the lower part of the body that are not captured, which is why they were excluded from the quantitative results of the tests, $j \notin \{13, 14, 15, 17, 18, 19\}$. Analyzing the results of the metrics for $EV1{:}3$, and despite the lightest backbones, $EV2{:}3$, presenting good results, the best ones were obtained in $EV1$, namely $EV1_5$, $EV1_7$ and $EV1_{10}$ with, 96%, 92% and 94% mAP, respectively. Moreover, $EV1_7$ is considered the best result due to its highest computational efficiency, $\frac{s/img(EV\#)}{s/img(EV\#)}$, (i.e. 745% for $EV1_5$ and 200% for $EV1_{10}$). Moreover, the difference between precisions can be seen from the different permutations: ($EV1_7$ to $EV1_5$) the reduced input image resolution and lack of refinement stages, reduce the spatial information and label refinement (i.e. confidence map and vector fields), respectively; ($EV1_7$ to $EV1_{10}$)

Table 4. Ablation and hyperparameter test results for CMU (EV1), MobileNetV1 (EV2), MobileNetV2 (EV3). The bold line represents the best results. Performance is assessed in mean accuracy (mAP), area under curve (AUC) and seconds per image (s/img). Bold line represents best results for each backbone evaluation.

EV1	**mAP**	**AUC**	**s/img**	**EV2**	**mAP**	**AUC**	**s/img**	**EV3**	**mAP**	**AUC**	**s/img**
$EV1_1$	92%	74%	0.0509	$EV2_1$	88%	71%	0.0249	$EV3_1$	**90%**	**73%**	**0.0189**
$EV1_2$	83%	66%	0.0499	$EV2_2$	89%	74%	0.0459	$EV3_2$	90%	74%	0.0209
$EV1_3$	83%	66%	0.0509	$EV2_3$	90%	75%	0.1257	$EV3_3$	90%	73%	0.0269
$EV1_4$	90%	71%	0.0808	$EV2_4$	90%	75%	0.04	$EV3_4$	91%	75%	0.024
$EV1_5$	96%	75%	0.208	$EV2_5$	90%	74%	0.0376	$EV3_5$	91%	75%	0.027
$EV1_6$	91%	76%	0.0369	$EV2_6$	**89%**	**72%**	**0.0247**	$EV3_6$	89%	72%	0.0209
$EV1_7$	**92%**	**72%**	**0.0279**	$EV2_7$	89%	72%	0.0269	$EV3_7$	90%	72%	0.0216
$EV1_8$	87%	70%	0.0499	$EV2_8$	88%	73%	0.0625	$EV3_8$	90%	74%	0.0269
$EV1_9$	91%	73%	0.0509								
$EV1_{10}$	94%	75%	0.0559								
$EV1_{11}$	91%	73%	0.0499								
$EV1_{12}$	93%	74%	0.0519								

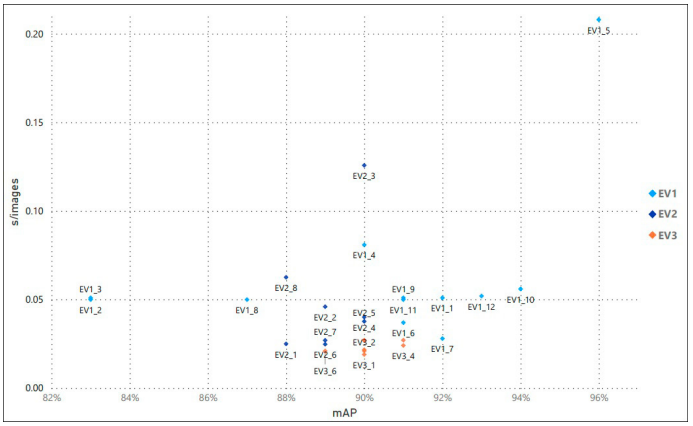

Fig. 4. Representative graph of mAP and s/img time metrics for each trial of each evaluation (EV1, EV2 and EV3)..

the increase of σ parameter from 8 to 10, allows a better convergence of the method, without significant loss of the localization ability. Considering $EV2$:3 results, $EV2_6$ and $EV3_1$ present the best results for each backbone configuration. Despite their reduced precision, comparatively to $EV1$ results, they present new oportunities for performance optimization (i.e. as shown in $EV2_6$ with its reduced number of filter calculations). As such, $EV3_1$ can be seen as the best algorithmic configuration, with 90%, 73% and 0.0189 of mAP, AUC and s/img respectively, thus reaching a better compromise of precision and performance, achieving 1100% of computational efficiency with a loss of 6% and 2% of mAP and AUC respectivey, compared to the highest precision model $EV1_5$. Qualitative results for this configuration are shown in Figure 5.

6. Conclusions and Future Work

In this paper, we show the customization of the OpenPose method for the estimation of the human body pose, to later use this kind of approach in the detection of human actions in SAVs.

To solve this task, we use of the OpenPose model with three different backbones (CMU, MobileNetV1 and MobileNetV2), ablation and hyperparameter configurations. The method that proved to be the best approach was the OpenPose method with: Depth based features; MobileNetV2 backbone; spatial and RGB augmentation disabled; 256x208 input resolution; $\sigma = 8$; Reduced LearningRate plauteau disabled; refinement stages removed (i.e. $T = 1$); feature map, F, output resolution 28x28. Moreover, reaching a mean Average Accuracy (mAP) of 90%, Area Under

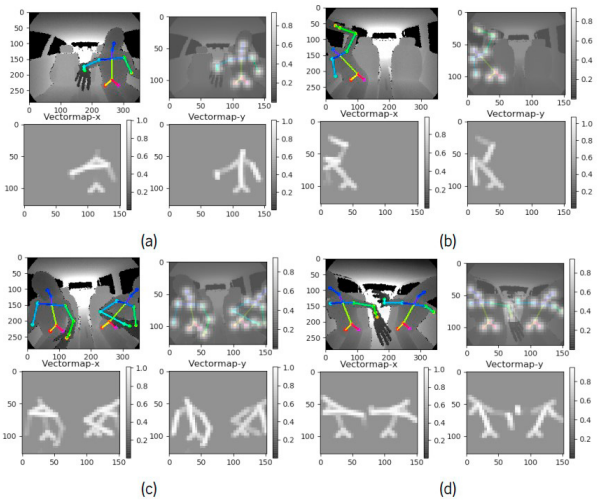

Fig. 5. Result of human pose inference on 4 different test images, with visualization of the estimated skeleton (top left graph), the predicted confidence maps (top right graph) and their respective connections (lower left and right graphs for X and Y coordinate coordinate, respectively).

Curve (AUC) of 73%, and 0.0189 seconds per image (s/img). From a qualitative point of view, it was observed that sometimes the trained models confused the left side of the body with the right side, and this may be associated with the depth images that are simpler and devoid of textures, making it more difficult for the trained model to interpret the captured image.

For future work a possible improvement could be to add links between joints in order to improve the network's ability to model spatial dependencies, and to test the algorithms on a real dataset.

Acknowledgements

This work is supported by: European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº 039334; Funding Reference: POCI-01-0247-FEDER-039334].

References

[1] Torres, H. R., Oliveira, B., Fonseca, J., Queirós, S., Borges, J., Rodrigues, N., Coelho, V., Pallauf, J., Brito, J., and Mendes, J. (2019). Real-Time Human Body Pose Estimation for In-Car Depth Images. In IFIP Advances in Information and Communication Technology, volume 553, pages 169–182. Springer New York LLC.

[2] Borges, J., Oliveira, B., Torres, H., Rodrigues, N., Queiros, S., Shiller, M., Coelho, V., Pallauf, J., Brito, J. H., Mendes, J., and Fonseca, J. C. (2020). Automated generation of synthetic in-car dataset for human body pose detection. In VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, volume 5, pages 550–557. SciTePress.

[3] Borges, J., Queiros, S., Oliveira, B., Torres, H., Rodrigues, N., Coelho, V., Pallauf, J., Brito, J. H. H., Mendes, J., and Fonseca, J. C. (2021). A system for the generation of in-car human body pose datasets. Machine Vision and Applications, 32(1):1–15.

[4] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014.

[5] W. Zhu et al., "Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks," 2016.

[6] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using LSTM and CNN," 2017 IEEE Int. Conf. Multimed. Expo Work. ICMEW 2017, pp. 585–590, 2017, doi: 10.1109/ICMEW.2017.8026287.

[7] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," Proc. - 2017 IEEE Winter Conf. Appl. Comput. Vision, WACV 2017, pp. 148–157, 2017, doi: 10.1109/WACV.2017.24.

[8] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 8, pp. 1963–1978, 2019, doi: 10.1109/TPAMI.2019.2896631.

[9] C. Caetano, F. Bremond, and W. R. Schwartz, "Skeleton image representation for 3d action recognition based on tree structure and reference joints," in Proceedings - 32nd Conference on Graphics, Patterns and Images, SIBGRAPI 2019, 2019, pp. 16–23, doi: 10.1109/SIBGRAPI.2019.00011.
[10] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in IJCAI International Joint Conference on Artificial Intelligence, Apr. 2018, vol. 2018-July, pp. 786–792, doi: 10.24963/ijcai.2018/109.
[11] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition with Shift Graph Convolutional Network." Accessed: Jun. 22, 2020. [Online]. Available: https://github.com/kchengiva/.
[12] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 3218–3226, 2015, doi: 10.1109/ICCV.2015.368.
[13] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, doi: 10.1109/CVPR.2018.00539.
[14] W. Du, Y. Wang, and Y. Qiao, "RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos," in Proceedings of the IEEE International Conference on Computer Vision, 2017, vol. 2017-Octob, doi: 10.1109/ICCV.2017.402.
[15] M. U. Khalid and J. Yu, "Multi-Modal Three-Stream Network for Action Recognition," in Proceedings - International Conference on Pattern Recognition, Sep. 2018, vol. 2018-Augus, pp. 3210–3215, doi: 10.1109/ICPR.2018.8546131.
[16] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, vol. 2019-June, pp. 12018–12027, doi: 10.1109/CVPR.2019.01230.
[17] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 1302–1310, 2017.
[18] ildoonet, "tf-pose," 2019. https://gitcode.net/mirrors/ildoonet/tf-pose-estimation.
[19] TensorFlow Core v2.7.0." [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/ReduceLROnPlateau. [Accessed: 11-Jan-2022].
[20] 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.
[21] 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.
[22] 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.
[23] 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.
[24] J. Shotton et al., "Real-Time Human Pose Recognition in Parts from Single Depth Images," Commun. acm, vol. 56, no. 1, 2013.
[25] 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.
[26] 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.
[27] 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.
[28] Z. Su, M. Ye, G. Zhang, L. Dai, and J. Sheng, "Cascade Feature Aggregation for Human Pose Estimation," Cvpr, 2019, [Online]. Available: http://arxiv.org/abs/1902.07837.
[29] 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.
[30] G. Borghi, "POSEidon: Face-from-Depth for Driver Pose Estimation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5494–5503.
[31] P. Murthy, O. Kovalenko, A. Elhayek, C. Gava, and D. Stricker, "3D Human Pose Tracking inside Car using Single RGB Spherical Camera," 2017.