IEEE Open Journal of
Intelligent Transportation
Systems

Received 4 July 2021; revised 7 November 2021; accepted 20 December 2021. Date of publication 6 January 2022; date of current version 2 February 2022.

Digital Object Identifier 10.1109/OJITS.2022.3140612

Automated Quantification of Occupant Posture and Shoulder Belt Fit Using Safety Specific Key Points

FRANZ HARTLEITNER[1,2], A. KOPPISETTY[2], AND K. BOHMAN[3]

[1]Computational Statistics and Data Analysis, Augsburg University, 86159 Augsburg, Germany

[2]Complete Vehicle Data Science, Volvo Car Corporation, 40531 Gothenburg, Sweden

[3]Safety Centre, Volvo Car Corporation, 40531 Gothenburg, Sweden

CORRESPONDING AUTHOR: F. HARTLEITNER (e-mail: franzhartleitner@fastmail.com)

ABSTRACT Virtual evaluation of automotive safety with variation in occupant posture and shoulder belt fit is gaining importance, and there is a need of methods facilitating analysis of occupant postures in driving studies. This study is aimed to develop an AI-based computer vision method to automatically quantify occupant posture and shoulder belt position over time in a car. Traceable defined key points on the occupant were related with the shoulder belt and quantified over time in real 3D coordinates by predefined key measurements, utilising the underlying spatial information of a Intel RealSense 3D Camera. The key points are defined as traceable key points relevant to relate the occupant to the vehicle environment and to estimate shoulder belt position. Key point prediction results suggest an average deviation of around 1cm per coordinate, which enable a reliable spatial categorization of the respective tracked occupant by analyzing the key measurements. This method providing continuous information of the occupant position and belt fit will be useful to identify common occupant postures as well as more extreme postures, to be used for expanding variations in postures for vehicle safety assessments.

INDEX TERMS Computer vision, neural networks, seating postures, transfer learning.

I. INTRODUCTION

OCCUPANT safety in vehicles is evaluated through crash tests, using anthropometric test devices (ATDs), which represent humans. Legal requirements and consumer rating crash safety programs are well described in test protocols. The ATDs are positioned in standardized sitting postures [1]. However, in real life, the variation of sitting postures may be larger than those represented by the standardized sitting postures used in current crash tests [2]. Sitting posture may vary depending on vehicle environment, personal preferences, anthropometric differences and vehicle dynamics. Sitting posture and belt fit may also vary over time, for comfort reasons. Consumer rating programs like EuroNCAP [3] and IIHS [4] have started to explore virtual testing besides traditional crash tests and that opens up for parameter studies including a greater range of sitting posture. There is a need to increase the knowledge of variation in sitting postures that takes place in real life. Therefore, reliable data collected in real life conditions is needed to verify the existing protocols or show potential for improvement.

In order to explore the 3D video footage data of occupant postures and belt fit in cars, it is often necessary to review single frames and document the information manually. This limits the analysis of driving studies to a subset of the whole data set, assuming it represents the whole trip. A method, which allows an automated analysis of videos would provide improved quality and allow to analyse all collected film data.

There are several methods to classify seat belt usage in cars, such as using 2D cameras inside the car. A portable warning system with data collected inside the car is proposed in [5]. The underlying detection method of the seat belt is a hard coded template, which utilizes the assumption, that the seat belt has an degree of around 45 degrees. The authors of [6] recognize incorrect positioning of the seat belt by the "acceptable distance of the seat belt from the neck" of drivers

The review of this article was arranged by Associate Editor Abdulla Hussein Al-Kaff.

or passengers to avoid neck injuries. They collect 4 different seat belt positions (fixed height of contact point of seat belt and b pillar) and classify, whether the seat belt fits in a proper way or not, dependent on the position of the height of the contact point. In [7], at first the shoulder and hip key points are detected and on the basis of that, the seat belt is detected with a feature vector to find the seat belt between the shoulder and hip key points. However, the method does not relate the shoulder belt position to the body key points. An approach to examine the seat belt fit of a person was suggested in [8], where a classifier separates various shoulder belt positions on a person. There are large data sets like Common Objects in Context (COCO) [9], which provide a large amount of key point annotations of persons in various positions. A notable attempt to process images, which utilizes the COCO data set, was proposed in [10]. Here the authors detected the seat belt with a semantic segmentation and conducted a human posture estimation by detecting the key points of a person. The mentioned data sets and methods in the literature focus on a 2D image evaluation and detect key points in a unreliable way, because the location of the key points are not explicitly defined in a traceable and repeatable manner.

There are several approaches to monitor sitting posture, and some use a categorization method. The authors of [11] collect 3D footage of persons sitting at a desk and classify the images into 14 different categories after applying feature enhancement. In [12] a quantitative assessment of posture with a wear-able monitoring system has been conducted to detect poor sitting postures causing neck or back pain. The system can give instant feedback to the user. In [13] a smart chair with pressure sensors is utilized to collect data. Afterwards the postures are classified into 15 different categories. A real time posture recognition system is proposed in [14], where a Kinect camera is used to collect point cloud data. Afterwards, the frames are classified into 8 different categories.

An interesting method to automatically analyze 3D videos was proposed in [15], which algorithmically tracks the head position of the passenger in a car by utilizing an underlying 3D point cloud data and partially analyses the shoulder belt positions manually. Reference [16] also monitored postures with a Kinect camera in order to classify different tasks of the driver with the underlying recorded information of head rotation vectors and upper body joints in 3D.

However, there are limited studies on real-time quantification of both posture and belt fit of occupants. Methods providing posture quantification over time can be used for several applications. Smart restraint systems can be adapted based on occupant posture and belt fit information, providing improved occupant protection. Furthermore, driver assistance application could benefit from this type of information. Reference [17] mentions, that driver state is a process over time, and by monitoring driver posture continuously it may be possible to predict posture changes the driver are about to do. A static image could not capture this temporal context, such as a continuous video stream. There is also a need to improve the knowledge of the variations of sitting postures, to include a greater variation of sitting postures when evaluating crash safety of cars.

FIGURE 1. Image examples of dataset X.

The aim of this study was to develop an end-to-end analytics pipeline for video data, which automatically quantifies the front seat passenger's sitting posture and shoulder belt position continuously in a car with key points, which are defined from an occupant safety and comfort perspective. This enables an automatic, continuous quantification of posture movement of passengers in cars by detecting key points and relating them with different restraint systems in a car.

II. METHODS

A. DATA COLLECTION

A driving study was conducted with 11 test persons (TP) as front seat passengers in a large passenger car. The number of participants was limited due to the Covid-19 situation. All test persons participated voluntarily and they consented to use of their video data for method development and for publishing. The total driving time took about 1 hour per passenger and included both rural and urban areas. Video frames were collected with an Intel RealSense Depth Camera D415 [18] with 3 frames per second, attached to the front window allowing a front view of the passenger. The camera captured the whole upper body, including the pelvis region. Technical details are denoted in the Appendix (Table 8), examples are visualized in Figure 1. A target marker was attached at the jugular notch, directly on the skin of the test person.

The camera supports a frame based video collection, which is able to capture various resolutions with color channels (RGB) and spacial channels (XYZ coordinates in meter) for each pixel as visualized in Figure 17. The camera measures the z coordinate, which is the depth distance of a pixel to the camera and computes the x (vertical position) and y (lateral position) coordinate with the provided Software Development Kit from Intel. Every frame of the recorded video is in the format $\mathbb{R}^{H\times W\times C}$: (Height H: 840, Width W: 480, Channel C: 6).

B. DATA ANNOTATION

Figure 2 shows an image with an orange bounding box, which defines the location of the person in the image with the COCO annotator [19]. Furthermore the seat belt is marked with a polygon.

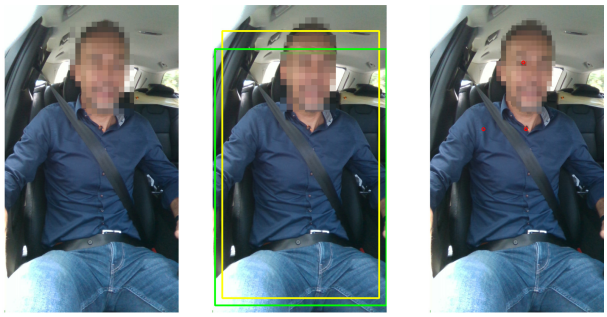
FIGURE 2. Example of one image in dataset X, annotated with COCO Annotator [19]; it supports bounding boxes, key points and polygons.

Several body key points were defined on the head, shoulder and upper sternum, and they were annotated according to a traceable, properly defined schema (see Figure 2).

The key point 'upper sternum' is defined as the surface point on the skin or clothing of a person at the jugular notch. The key point will be used to trace the torso x, y, z position of a person in relation to the center line of the seat. The jugular notch is a bony landmark easy to identify on a person, which increases the quality of the data collection. This landmark will give information if the person is moved laterally from the center line of the seat, but also if the person is leaning forward.

The two vertical torso lines are defined laterally alongside the torso through the armpits. Secondly, a horizontal line is defined at the level of the 'upper sternum' key point in parallel to the upper torso. The key points of the shoulders are defined by moving 1cm (in real coordinates) from the intersections of the horizontal upper sternum line and vertical torso lines to the center of the person. The key point 'right shoulder' and 'left shoulder' are related to the respective side from the occupant's perspective. The outboard shoulder (meaning right shoulder in this data collection) is of particular interest, because it will be relate the shoulder belt to the shoulder key point, providing information if the shoulder belt is off the shoulder or far out on the shoulder, which are safety critical shoulder belt positions. The 'eye' key points are defined as the center of the respective pupil and is used to quantify the head position. From safety perspective, it is relevant to understand how the head is positioned relative the head restraint.

Other key points like the nose, ears, elbows and wrists are also present in the dataset. They are more straightforward to define and therefore an explicit definition has been omitted, an example annotation is visualized in Figure 2.



(a) Image example of dataset **X**.

(b) Annotated (yellow) and predicted BB (green).

(c) Example of a key point prediction of the 'HR-Net refined'.

FIGURE 3. Visualization of pipeline with and without ground truth bounding boxes.

C. SEMANTIC SEGMENTATION

A ResNet-101 backbone was used in combination with a feature pyramid network. The goal was to obtain a binary pixel wise classifier for the two classes 'seat belt' and 'background'. The training process was conducted with the 271 training, 68 validation and 223 test samples. The sampling process of the training and validation data, which is an heuristic approach to ensure a high variance within the data, is outlined in Appendix-D. The test dataset was obtained by sampling every 500^{th} frame from dataset **X**. Each image has a corresponding ground truth annotation, which defines, if a pixel belongs to the class 'seat belt' or not. The model architecture of the ResNet-FPN allows an image input of the dimension 224x224x3. For dataset *X* the whole image was resized to the desired dimension and used as input. The processing pipeline of the semantic segmentation is illustrated in Figure 12. At first, the original image was resized. Then the network passed the input forward and assigned a class to each pixel. Resizing the prediction yielded the prediction with the same size as the input image. Details of the training process are denoted in the Appendix.

D. HUMAN POSTURE ESTIMATION

On the key point leader board of the COCO website [9] one can see the currently best performing neural network architectures, which detect the key points of a non-published test set. The HR-Net achieves very good results on this benchmark. The MIT licence and a good code quality with a well maintained GitHub repository, to be able to further extend the method, led to the selection of the HR-Net [20]. The pipeline is depicted in Figure 3. The original input in Figure 3(a) is the basis for the detection of the person in the image. For the 'refinement' of the HR-Net, which means the re-training and validation of the HR-Net with self-annotated data, ground truth bounding boxes (Figure 3(b)) of the person with corresponding key points were provided. These bounding boxes

were cropped, slightly adjusted and then used as input for the HR-Net refined, a prediction of relevant key points is visualized in Figure 3(c). The training procedure, including bounding box preparation and data augmentation of [20] to predict the 18 key points with relocated COCO key points as defined in Section II-B and Upper Sternum) was replicated.

The focus of the key point detection task was to accurately detect the key points within the given environment. This was done by adjusting the last layer from the dimensions 96x72x17 to 96x72x18 and re-training the network as described below.

1) LOCATING THE PERSON IN AN IMAGE

The key points will be predicted with detected bounding boxes of persons, obtained by the neural network 'Faster R-CNN ResNet-50 FPN' [21]. The quality of the key point detections on unseen data depends on the accuracy in detection of the person. Thus, to measure the 'real' performance of the pipeline, the person detection results of 'Faster R-CNN ResNet-50 FPN' were used as a basis to measure the performance on unseen data in the testing stage.

For the scope of this work it would have been ideal, if the bounding boxes of the persons, which were detected by the 'Faster R-CNN ResNet-50 FPN', would have been detected with a high confidence. This could be realized, as each of the region proposals (bounding boxes) of the 'Faster R-CNN ResNet-50 FPN' come with 'scores that estimate probability of object or not object for each proposal' [21]. The score is a value between 0 and 1, in the context of this work 0 means the region proposal contains a person with low probability, and 1 that there is a person in the image. The 'acceptance'-threshold for the present pipeline was set to 0.9, in order to only detect regions where a person is in the image with a very high confidence. Furthermore, only region proposals with a reasonable size of a width of more than 150 pixels were accepted, excluding very small bounding boxes which do not contain a person. Further reasoning regarding the relation between locating the person in the image and the key point prediction is attached in the Appendix.

2) 2D-ACCURACY HUMAN POSTURE ESTIMATION

In order to measure the accuracies of human posture estimation in 2D, key point similarity (KS), mean key point similarity (mKS) and object key point similarity (OKS) were defined. The following definitions elaborate how these metrices are mathematically calculated.

Definition 1: The metric Key Point Similarity (KS) measures the similarity between the prediction and the ground truth of one key point. It is defined as

$$\mathrm{KS}(\mathrm{kpid}) = \exp\left(\frac{-\left\|\hat{kp}_{kpid} - kp_{kpid}\right\|^2}{2s^2k_{kpid}^2}\right), \quad kpid \in \{0,\ldots,17\}$$

with the negative squared Euclidean distance between the ground truth location of key point *kpid*, denoted as kp_{kpid} and the predicted location of key point *kpid*, denoted as $\hat{kp}_{kpid}$, the object scale s (root of the area of bbp_j) and a key point dependent constant k_{kpid} which controls the decline of the Gaussian function.

Definition 2: The metric Mean Key Point Similarity (mKS) measures the similarity between the prediction and ground truth of one key point over the whole dataset X, which includes the bounding boxes bbp_j, $j = 1,\ldots,N$, which define the location of the persons. The mKS is defined as

$$\mathrm{mKS}(\mathbf{X},\mathrm{kpid}) = \frac{1}{N}\sum_{j=1}^{N}\exp\left(\frac{-\left\|\hat{kp}_{kpid,j} - kp_{kpid,j}\right\|^2}{2s_j^2k_{kpid}^2}\right), \quad kpid \in \{0,\ldots,17\}$$

with the negative squared Euclidean distance between the ground truth location of key point *kpid*, denoted as kp_{kpid}, and the predicted location of key point *kpid*, the object scale s_j (root of the area of bbp_j) and a key point dependent constant k_{kpid} (numerical values can be found in Table 7) which controls the decline of the Gaussian function.

Furthermore, the metrics Object Key Point Similarity (OKS) and Average Precision were used as defined in the COCO challenge [9].

3) 3D-ACCURACY HUMAN POSTURE ESTIMATION

Besides hardware related constraints, predictions naturally deviate from ground truth annotations. The metric *average 3D-accuracy* measures the deviation of one key point on coordinate level. The following definition can be used interchangeably for the x, y and z coordinate:

Definition 3: Let $bbp_j, j \in \{1,\ldots,N\}$ be ground truth BB annotations, which define the location of a person in dataset $\mathbf{X}$. Furthermore, let $kp_{kpid,j,x}$ be the annotated, ground truth x-coordinate of a key point and $\hat{kp}_{kpid,j,x}$ be the corresponding prediction of the key point of bbp_j. The average accuracy of the x-coordinate in 3D (aa_3D_x) for the whole dataset is determined by

$$aa\_3D_{x,kpid} = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\left(kp_{kpid,j,x} - \hat{kp}_{kpid,j,x}\right)^2}$$

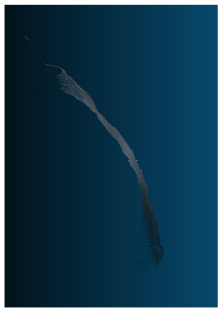
The calculation is possible for each key point, which is defined by the '*kpid*' from Table 7.

Furthermore, the *absolute average deviation* for one key point will be provided.

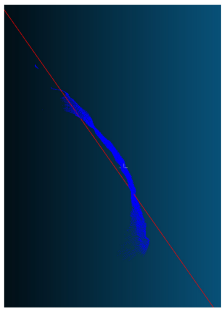
Definition 4: Let $bbp_j, j \in \{1,\ldots,N\}$ be ground truth BB annotations which define the location of a person in dataset $\mathbf{X}$. Furthermore, let $kp_{kpid,j,x}$ be the annotated, ground truth x-coordinate of a key point and $\hat{kp}_{kpid,j,x}$ be the corresponding prediction of the key point of bbp_j. The absolute average accuracy in 3D (aaa_3D) for the whole dataset is determined by

$$aaa\_3D_{abs,kpid} = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\sum_{i\in\{x,y,z\}}\left(kp_{kpid,j,i} - \hat{kp}_{kpid,j,i}\right)^2}.$$

(a) Illustration of seat belt pixels, which have been used to fit a 3D line (green) and excluded (red) to approximate the seat belt.



(b) Illustration of pixels classified as 'shoulder belt' in 3D space (visualized with software CloudCompare [22], pelvis belt is excluded).



(c) Illustration of line fit (red). Line has been fitted to pixels classified as 'shoulder belt' (blue). Visualized with CloudCompare [22]. Pelvis belt is excluded.

FIGURE 4. Pipeline to approximate the shoulder belt with a line. Seat belt pixels which capture the torso are selected, then a line is fitted through the spacial information (xyz-coordinates) of these pixels.

The calculation is possible for each key point, which is defined by the '*kpid*' from Table 7.

E. KEY MEASUREMENTS

1) SHOULDER BELT APPROXIMATION

The shoulder belt was approximated with a line, to facilitate the computation of the distances related to the shoulder belt. The seat belt consists of a shoulder belt restraining the torso and a lap belt restraining the pelvis. In order to obtain a good line fit, which captures the shape of the shoulder belt, the segmentation of the lap belt must be neglected. Since all the videos were captured from the same angle and the lap belt was approximately at the same position for all test persons in the data, all the pixels below the 600^{th} row in the image were treated as 'back ground'. Figure 4(a) shows the pixels (green), which were used to compute the line fit, and those which were excluded (red).

All available spacial information of the seat belt pixels, which were above the 600^{th} row, were used as a basis for the further computation, depicted in Figure 4(b). Note here, that the camera does not capture spacial information of all pixels (see Figure 17(a)).

Let $A \in \mathbb{R}^{N\times 3}$ be the points in three dimensional space and $\bar{A}$ be the centered points around the column wise mean x_0 of the data. The best line fit is defined as the line

$$l = x_0 + \lambda x, \;\; \lambda \in \mathbb{R}, \tag{1}$$

which minimizes the squared perpendicular distances d between the line l and the centered points $\bar{A}$,

$$\sum_{i=1}^{N} d(a_i, l)^2.$$

Utilizing NumPy's [23] singular value decomposition, the left-singular vectors, the eigenvalues and the right-singular vectors were computed. The right-singular vector corresponding to the largest eigenvalue of $\bar{A}$ yields the desired vector x in (1), which minimizes the sum of the squared perpendicular distances of the points $\bar{A}$ to the line l [24]. An example is visualized in Figure 3(c).

2) SHOULDER BELT DISTANCE TO RIGHT SHOULDER

The first key measurement is the shortest distance of the seat belt line to the key point 'right shoulder'. The calculations were conducted with the approximated seat belt line l and the spacial information of the key point right shoulder, which are denoted as $kp_{rs,xyz}$. The shortest distance is the length of the perpendicular line between the 3D-line l and the 3D key point $kp_{rs,xyz}$ and was computed according to the following steps:

1) Calculate centered key point $\bar{kp}_{rs,xyz} = kp_{rs,xyz} - x_0$
2) Project the centered key point on to the direction vector x of line l: $p = \frac{<x,\bar{kp}_{rs,xyz}>}{\|x\|} \cdot x$
3) Calculate the distance $d_{rs,sb}$ between right shoulder (rs) and seatbelt (sb)
   $d_{rs,sb} = d(\bar{kp}_{rs,xyz}, p) = \|\bar{kp}_{rs,xyz} - p\|$
4) Determine sign s, whether the y-component of the key point is above or below the projection:
$$s = \begin{cases} -1, & \text{if } \bar{kp}_{rs,y} < p_y \\ 1, & \text{if } \bar{kp}_{rs,y} > p_y \end{cases}$$
5) The result is $seatbelt\_dist = s \cdot d_{rs,sb}$

The y-axis was taken as a basis to determine, if the key point 'right shoulder' is above or below the approximated line of the shoulder belt. This attempt turned out to be unreliable, likely due to a slightly skewed camera angle and therefore a skewed 3D-coordinate system. Therefore the y-axis in 2 dimensional space (i.e., the pixel coordinate system) was taken as a basis to determine the sign, whether the shoulder key point was below (−) or above (+) the approximated

TABLE 1. The metric intersection over union (*IoU*) for training, validation and test datasets; evaluation has been done by the best model state according to the validation dataset.

	IoU training	IoU validation	IoU test
dataset $\mathbf{X}$ WM	0.766	0.773	0.750

shoulder belt line in step 4. Figure 14 illustrates the conducted computation on a 2D, pixel-wise basis. Note here, that for the visualization purpose, the shoulder belt approximation (blue line) and the projection of the key point 'right shoulder' onto the line (yellow point) in Figure 14 have been calculated on a pixel wise (2D) level separately.

3) UPPER TORSO POSITION

The upper torso position is defined with the spacial information of the predicted key point 'upper sternum', which will be denoted as $kp_{us,xyz}$.

To ease the interpretation of the measurement, it was centered around a point q_0 on the seat, which was defined and illustrated in the Appendix-C in Figure 18. This center q_0 could be different for each test person, due to seat adjustments prior the drive started. No adjustments of the seats were done during the ride. Therefore, the final configuration of the seat (Figure 18) after the drive was used for each test person to calculate the measurement. The result of the torso position of the person of interest is a centered key point $\bar{kp}_{us,xyz} = kp_{us,xyz} - q_0$ for each frame, which measures the lateral, vertical and depth position between the point on the seat and the key point upper sternum.

4) POSITION OF THE HEAD IN RELATION TO THE HEAD REST

Another key measurement is the position of the head in relation to the head rest. It was calculated with the spacial information of the key point 'right eye' (denoted as $kp_{right\_eye,xyz}$), which was set in relation to a fixed point on the center-line of the headrest x_0 (Appendix, Figure 19):

$$\bar{kp}_{left\_eye,xyz} = kp_{left\_eye,xyz} - x_0.$$

The determination of this fixed point were done for all persons after they left the car, in order to take the correct seat configuration into account.

III. RESULTS

A. SEMANTIC SEGMENTATION

The 2D result of the training process is reported in Table 1 in the metric Intersection over Union (IoU). The IoU of the test set is 0.75. A comparable prediction quality is depicted in the Appendix in Figure 12, which shows a prediction of the test set with an IoU of 0.7.

B. 2D ACCURACY KEY POINTS

1) RESULTS WITH GROUND TRUTH BOUNDING BOXES

Table 2 reports the mKS on test dataset X, computed with ground truth bounding boxes. Results are provided for the original HR-Net and the refined version (HR-Net refined).

TABLE 2. Results of human posture estimation on test dataset X. Mean key point similarity (mKS) is provided for the HR-Net and HR-Net refined. The numbers in this figure consider the case, when the test dataset is defined by annotated ground truth bounding boxes.

Key point	HR-Net	HR-Net refined	Tot. ann.
right eye	0.9780	0.9751	222
right shoulder	0.9755	0.9937	222
upper sternum	n.a.	0.9956	222

TABLE 3. Results of human posture estimation on test dataset X. Mean Key point similarity (mKS) is provided for the HR-Net and HR-Net refined. The numbers in this figure consider the case, when the test dataset is defined by bounding boxes, which are obtained by predictions of the "Faster R-CNN ResNet-50 FPN" [21].

Key point	HR-Net	HR-Net refined	Tot. ann.
right eye	0.9506	0.9512	183
right shoulder	0.9637	0.9765	183
upper sternum	n.a.	0.9794	183

The prediction quality of the right eye was about the same for both networks, around 0.97 mKS. The right shoulder averaged at 0.9755 mKS for the HR-Net and at 0.9937 mKS for the HR-Net refined. The Upper Sternum had a mKS of 0.9956 for the refined version and is not detected by the original architecture. Figure 10 visualizes these numerical values. 'Tot. ann.' is the amount of ground truth annotations, which are present (visible) in the test dataset of $\mathbf{X}$.

2) RESULTS WITH REGION PROPOSALS OF A PERSON DETECTOR

The mKS will be provided for all the detected bounding boxes $\hat{bbp}_j, j \in \{1, \ldots, N\}$ by the 'Faster R-CNN ResNet-50 FPN' network. Note here, that these results were obtained by evaluating the detected bounding boxes (183) and neglecting the rest of the ground truth bounding boxes (39), which were not detected.

The right eye has a mKS of around 0.95 for both networks (Table 3). The key point right shoulder averaged at 0.9636 (HR-Net) and 0.9765 (HR-Net refined) respectively and the Upper Sternum was detected with an mKS of 0.9794 by the HR-Net refined. These numerical values are visualized in Figure 11.

C. 3D ACCURACY KEY POINTS

1) 3D RESULTS WITH REGION PROPOSALS OF A PERSON DETECTOR

Table 4 depicts the results for test dataset $\mathbf{X}$, where the predictions of the HR-Net refined were based upon bounding boxes, which were predicted by the 'Faster R-CNN ResNet-50 FPN' [21]. The coordinate-wise average accuracy were at around 1cm for each key point, the absolute deviation range was around 1.5-2cm. There were 183 bounding boxes, which define the location of persons. As described above, the HR-Net refined predicted a 2D pixel location in an image for each key point. The 3D spacial information (x, y and z coordinates) at these 2D pixel locations were compared with the ground truth spacial information in Table 4. There are few missing 3D measurements when comparing 'Amount of

FIGURE 5. Illustration key points in a 3D point cloud.

TABLE 4. 3D accuracy results of human posture estimation with predicted bounding boxes by the "Faster R-CNN ResNet-50 FPN" [21] on test dataset X in cm. The 3D average accuracy for the x, y and z coordinate and the absolute accuracy is provided.

Key point	aa_3D_x	aa_3D_y	aa_3D_z	aaa_3D_{abs}	AoM
right eye	1.3254	0.7185	0.5287	1.5977	181
right shoulder	1.3976	1.1613	0.7480	1.9651	181
upper sternum	0.9440	0.6773	0.5261	1.2754	180

TABLE 5. 3D accuracy Results of human posture estimation with annotated, ground truth bounding boxes on test dataset X in cm. The 3D average accuracy for the x, y and z coordinate and the absolute accuracy is provided.

Key point	aa_3D_x	aa_3D_y	aa_3D_z	aaa_3D_{abs}	AoM
right eye	0.4750	0.4843	0.4067	0.7910	222
right shoulder	0.7186	0.6304	0.5818	1.1190	220
upper sternum	0.5695	0.3952	0.3788	0.7900	217

Measurements' (AoM) in Table 4 with 'Tot. ann.' in Table 3. These missing 3D measurements originate from 2D key point locations which did not contain spacial information, due to restrictive camera properties (Figure 17(a)), hence they were neglected.

2) 3D RESULTS WITH GROUND TRUTH BOUNDING BOXES

Table 5 depicts the results for test dataset **X**, where the predictions of the HR-Net refined were based upon annotated ground truth bounding boxes. There were 222 bounding boxes, which defined the location of a person. The coordinate-wise average accuracy was between 0.4 cm and 1 cm and the absolute accuracy range was around 1cm.

D. 3D RESULTS KEY MEASUREMENTS

1) RESULTS SHOULDER BELT—SHOULDER DISTANCE

Empirically measured results on the basis of one test drive with TP 16 are illustrated in Figure 7. A negative value (−) of the measurement means the key point 'right shoulder' is below, a positive (+) that the key point is above the shoulder belt. In the case of TP 16, distances of around -15cm, as depicted in Figure 14(a), can be regarded a good fit of the shoulder belt. The frames with distances above 0 indicate shoulder belt positions on the outboard side of the shoulder

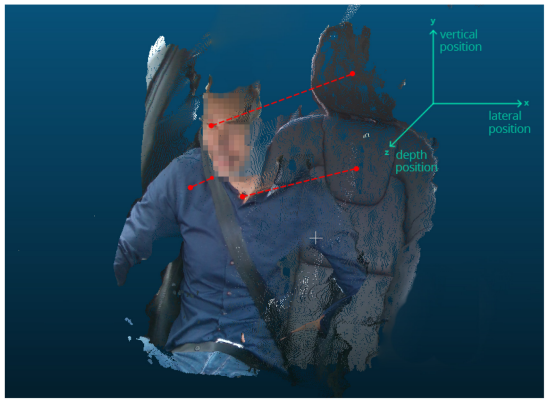


FIGURE 6. Illustration of the key measurements "shoulder belt to shoulder distance," "torso position" and "head position".

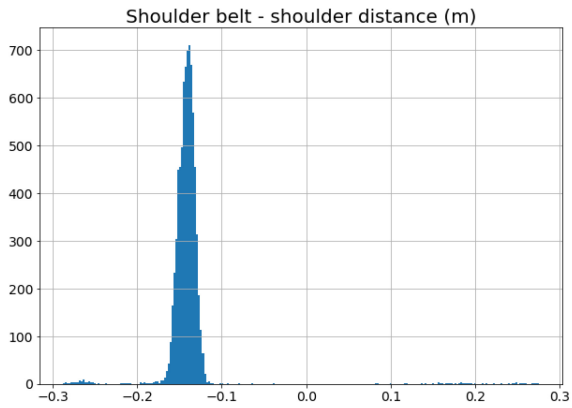


FIGURE 7. The key measurement "shoulder belt distance to right shoulder" of TP 16.

key point, with one example shoulder belt off the shoulder as seen in Figure 14(b).

2) RESULTS TORSO POSITION

The lateral positions of the upper sternum (Figure 8), indicate a tendency inboard skewness compared to the centerline of the seat. The centerlized values of the depth position indicate limited movements of extensive forward leaning of the upper torso.

3) RESULTS HEAD POSITION RELATED WITH HEAD REST

The head position relative the head rest, shows a spread in lateral position with a higher tendency of inboard position relative the centerline of the seat (Figure 9). There is limited excessive forward position of the head.

IV. DISCUSSION

In order to optimise for the highest possible accuracy within the key point detection task, the same persons are contained in the training, validation and test set, as a very good generalization could not be expected due to the difficulty of the task and the limited amount of test persons.

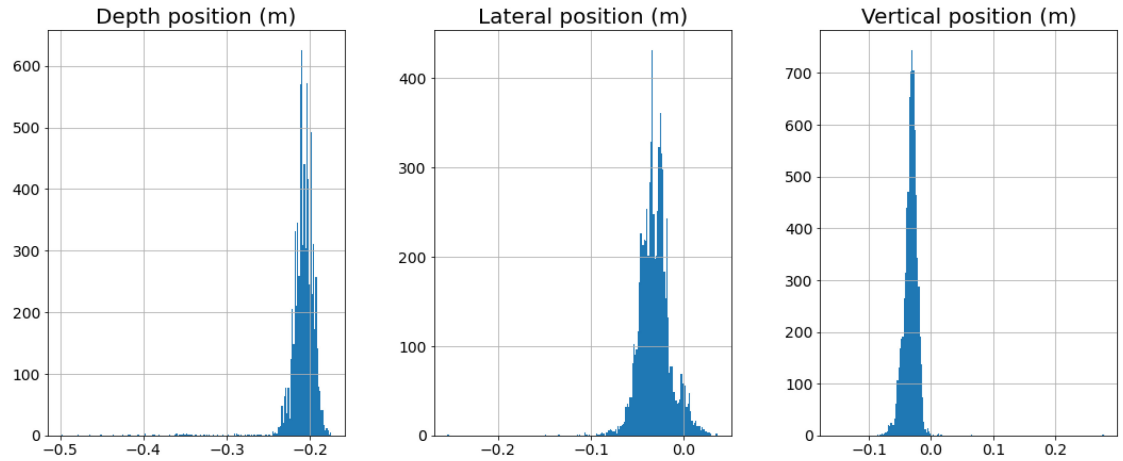


FIGURE 8. Histograms of TP 16, which depict the vertical (x-axis), lateral (y-axis) and depth position (z-axis) as distance between the key point upper sternum and a point q_0 on the seat.

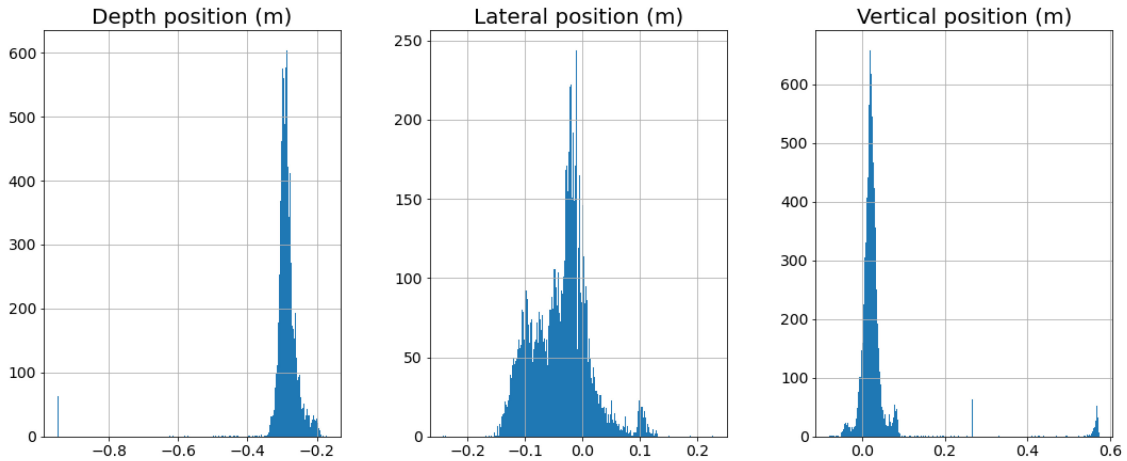


FIGURE 9. Histograms of the vertical (x-axis), lateral (y-axis) and depth position (z-axis) of the head in relation to the headrest, for TP 16.

Intel guarantees the z-accuracy (depth position) of the camera to be within a 2 percent range, the distance between the occupant and the camera in the present setting is around 1m. However, the provided results suggest a better accuracy.

A. ACCURACY SHOULDER BELT APPROXIMATION

The 2D-segmentation results of the shoulder belt (Table 1) clearly captured the shape of the shoulder belt (Figure 12). As the underlying spacial information of the 2D pixels are used to approximate the shoulder belt (Section II-E), the 3D approximation can assumed to be accurate.

B. HUMAN POSTURE ESTIMATION

1) 2D RESULTS

The original HR-Net architecture does not detect the key point 'upper sternum' and the shoulder key points are not well defined (Tables 2, 3). Therefore, there was a need of a precise, traceable definition of the key points in Section II-B and retraining the original network with an additional key point 'upper sternum' and refined shoulder positions. The stated results in Table 2 compare the performance of the networks on this definition of the key points, when the location of the person in the image was annotated by a ground truth bounding box. A visual comparison of the average network performance is depicted in Figure 10, where one can see a significantly better average performance of the 'HR-Net refined'. The general performance of both networks drop, when the ground truth bounding boxes, which defined the location of a person in the image, were replaced by predicted bounding boxes by the 'Faster R-CNN ResNet-50 FPN', as denoted in Table 3 and visualized in Figure 11. A large contribution of this drop in average accuracy originates from wrongly detected bounding box predictions. When a bounding box without a person was used as input for the HR-net, the network predicted a location for each key point anyway. In the present test set, 3 out of 183 bounding boxes do not contain a person. Fortunately they can be filtered and neglected in a later stage: Certain key measurements, which are computed with these key points, deviate around a magnitude from the desired key measurements based on the predicted bounding boxes.

(a) Illustration of mKS results of the original 'HR-Net' architecture from Table II.

(b) Illustration of mKS results of the 'HR-Net refined' from Table II.

FIGURE 10. Comparison of predictions of original and refined HR-Net. The key point upper sternum is not present in the original network (Figure 10(a)) and the shoulder key point is not located as defined in Section II-B.

(a) Illustration of mKS results of the original 'HR-Net' architecture from Table III.

(b) Illustration of mKS results of the 'HR-Net refined' from Table III.

FIGURE 11. Comparison of predictions of original and refined HR-Net. The key point upper sternum is not present in the original network (Figure 11(a)) and the shoulder key point is not located as defined in Section II-B.

C. 3D RESULTS

The results of most key points in Table 5 significantly improve in comparison to Table 4, and thus outline the potential of increasing average accuracy by improving the quality of the bounding boxes. However, the results in Table 4 with an average per coordinate accuracy deviation of approximately 1cm per key point constitute a reasonable basis for the further computations.

D. KEY MEASUREMENTS

The findings of the present work were encouraging, as it enabled a detailed data analysis of the collected video footage. It was possible to quantitatively summarize the

- shoulder belt to shoulder key point
- torso position
- head position

in a reliable way. A visualization of the key measurements is provided in Figure 6.

An example is depicted in Figure 14(a), where one can see a good shoulder belt fit, where the shoulder belt is located on the inboard side of the shoulder key point, indicating a good shoulder belt fit with shoulder belt positioned on the mid shoulder. In Figure 14(b), the shoulder belt is positioned under the arm, a shoulder belt fit deviating from the norm. The key measurement shoulder belt to shoulder key point indicates a positive value as the belt moves off the shoulder. A shoulder belt fit which deviates from the norm is illustrated in Figure 14(b). It depicts the setting where the shoulder key point is above the shoulder belt, which always is the case in the histogram range with positive distances. The continuous nature of the distances enable a more detailed analysis than proposed shoulder belt classifiers, which classify certain shoulder belt fits like 'belted', 'unbelted',. . ., 'under shoulder' in [8]. By filtering specific intervals within the histograms of the key measurements, qualitative results depicted in Section II-E were obtained.

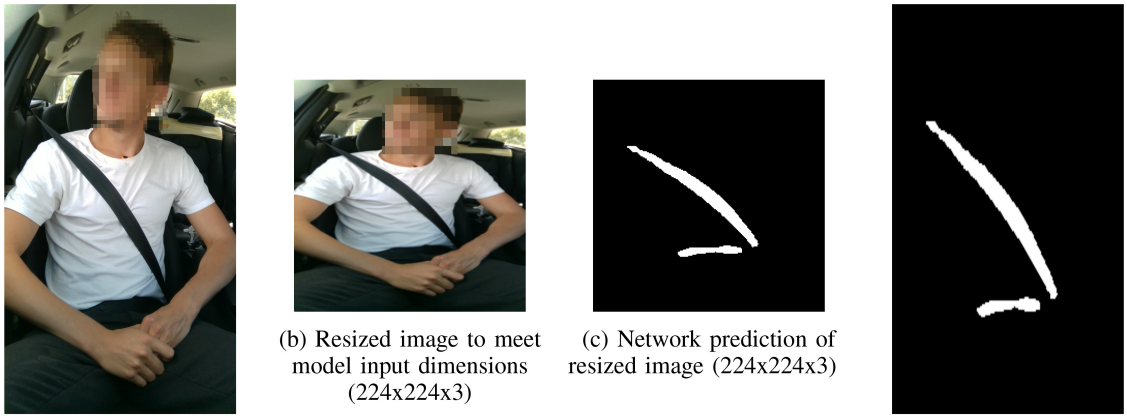
The key measurement "position of the head in relation to the head rest" shows whether the head is in line with the head rest (lateral position in Figure 9) and also quantified the depth distance between the head and the head rest (depth position in Figure 9).

Similarly, the key measurement 'Upper Torso position' in Figure 8 indicates, whether the torso is in a centralized position or not. In a minor fraction of this key measurement, the evaluation was not reliable. Some predictions, where no marker was visible at the upper sternum, seemed to be worse than the testing results from Section III-B2 suggested. A possible explanation is an overfitting behavior of the neural network to the distinctive marker.

The quantitative analysis of the detection results are also worth to note. Dependently on the TP, a fraction of around 40-90 percent of the frames were captured within the mentioned histograms above. The percentage of the respective percentage for each person can be found in Table 6.

There are several explanations, why some frames were neglected in the analysis, including no detected person on the image, the detected region proposal for the person does not fit criteria or no segmented shoulder belt on the image. Furthermore, in the beginning or the end of a test drive, there were no TP seated in the car.

Therefore, it was necessary to only accept region proposals of persons which were detected with a high confidence and further filtering them, in order to have a good basis for the key point detection in the next step. This trade off accounted for the majority of the neglected frames. This fraction of processed frames can be increased by decreasing the confidence of the region proposals. However, this would lower the average performance in Table 4 and Table 3.



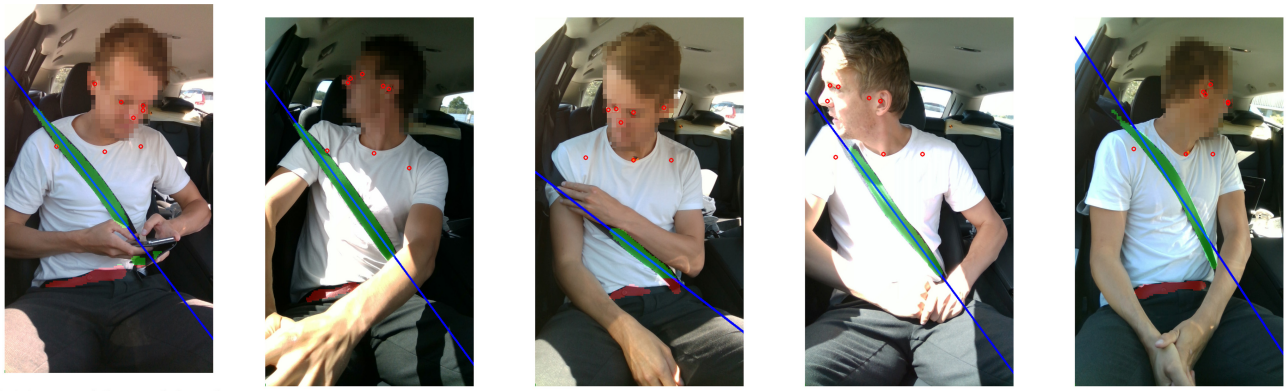
(a) Original image of dataset (848x480x3)

(b) Resized image to meet model input dimensions (224x224x3)

(c) Network prediction of resized image (224x224x3)

(d) Resized network prediction (848x480x3)

FIGURE 12. Pipeline for segmentation in order to detect the seat belt and background pixels. The predicted mask in sub figure (d) has an *IoU* of around 0.7.



(a) Forward leaned head - distance to head rest is large

(b) Seat belt is below shoulder key point

(c) Very flat 3D seat belt angle

(d) Torso leaned to right hand side

(e) Head leaned towards left hand side

FIGURE 13. Illustration of extreme postures, which can be samples by filtering for values at the tail of the respective Histograms.

Future Work: The shoulder belt was approximated by a line. When the shoulder belt was fitted to a person, it usually incorporated a curvature around the torso of a person. Other shapes than a line might be better to capture this curvature, and thus need further investigation.

The detection of the person by the 'Faster R-CNN ResNet-50 FPN' [21] in the pipeline constituted the basis for further computations and thus was important for the overall performance. As it was not focused on this part, an easy performance gain might be accomplishable by fine tuning the hyper-parameters in the existing pipeline or experimenting with new models in order to detect persons. The potential improvement of this step can be revealed by comparing Table 4 with Table 5. Detecting the upper sternum in a more robust way is likely be solvable by conducting driving studies without the corresponding marker and train the network on data without marker. The expected outcome would be a prediction, which is less dependent on the marker, thus comparable results to the shoulder key points can be assumed.

By including acceleration data to the analysis, the occupant movement can be connected to the vehicle dynamics, in order to understand if the movement is driven voluntarily by the occupant themselves or by vehicle dynamics. To accomplish a more complete quantification of the occupant posture, it would be desirable to also track the hip, to understand if the occupant is slouching forward with the hip. Furthermore, in order to quantify leg and feet position, an additional camera view would be necessary.

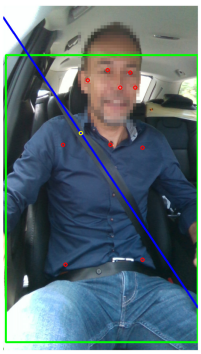
V. CONCLUSION

With the underlying spatial information of the Intel RealSense 3D camera [18], the shape of the shoulder belt was approximated by a line and the spatial information of the occupant key points were utilized to define occupant position relative to the vehicle and also shoulder belt fit. This method, providing continuous information of the occupant position and shoulder belt fit, will be useful to identify common occupant postures as well as more extreme postures, to be used for expanding variations in postures for vehicle safety assessments.

APPENDIX

A. PIPELINE SEMANTIC SEGMENTATION

See Figs. 12 and 13.



(a) Illustration of seat belt line fit (blue line), key points (red dots) and projection of right shoulder onto line (yellow).

(b) Qualitative example of the key measurement 'seat belt distance to right shoulder', which depicts the case where the seat belt is below the shoulder.

FIGURE 14. Key measurement "seat belt distance to right shoulder." The left image (14(a)) depicts the case, where the key point "right shoulder" is below (−) the seat belt. The image on the right hand side (14(b)) shows an image where the key point is above (+) the seat belt.

(a) Illustration of the key point 'upper sternum' (red), which is the basis of the key measurement 'torso position.'

(b) Qualitative example of the key measurement 'torso position', where the person is leaning to the left hand side.

FIGURE 15. Definition and qualitative example of the key measurement "torso position".

B. QUALITATIVE EXAMPLES OF KEY MEASUREMENTS

See Figs. 14–16.

C. SUPPLEMENTARY MATERIAL

See Fig. 17.

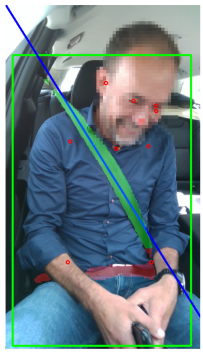
D. DATA SAMPLING

The whole dataset with the resolution 848x480x6 consisted of 110994 frames, stored within 13 different files. Two of the files were recorded for test purposes. The remaining 11 files respectively correspond to a test drive with a test person (tp):

The aim of the following sampling process was to reduce the number of frames and ensure a high variance in the dataset, in order to train and validate the neural networks in this work.

The following steps were done separately for all frames of each test person:

(a) Visualization of the key measurement 'position of the head in relation to the head rest'.



(b) Qualitative example of the key measurement 'position of the head in relation to the head rest'. The image is sampled by filtering the corresponding histogram 'Depth position'.

FIGURE 16. Definition and qualitative example of the key measurement "position of the head in relation to the head rest".

TABLE 6. Table, which shows the amount of processed frames and the total amount of frames by the pipeline per test person. Furthermore, the fraction of the processed frames has been computed.

test person	processed frames	total frames	percentage
TP 11	8260	9351	0.883
TP 14	4296	9064	0.473
TP 8	6037	9696	0.622
TP 15	6607	10404	0.635
TP 18	7273	9734	0.747
TP 19	4778	9862	0.484
TP 20	8972	9280	0.966
TP 17	3896	9781	0.398
TP 16	7755	9199	0.843
TP 12	6527	10538	0.619
TP 13	11161	12152	0.918

1) down-sampling the images (rgb information 848x480x3) to gray scale (64x48x1)
2) performing a principal component analysis of all gray scale frames of one person
3) clustering all frames of one person with PCA-components.

The following Figure depicts the PCA components of tp 16:

The rule for the pipeline was to include as many principle components as needed to explain at least 75 percent of the variance. The study of Figure 22 revealed, that around 20 PCA components were necessary for tp 16.

On the PCA features created above, a k-Means clustering was performed with different numbers of centroids. The associated cost can be seen below:

A similar, almost power like decent without clear cut-off for all test persons as illustrated in 22 was observed, a rule was set to include 30 centroids for each file. These

TABLE 7. Output channels of the HR-Net are linked by the key point ID (*kpid*) with the key point names. Each *kpid* corresponds to the respective channel in the last layer of the HR-Net. Key point loss weight ρ_{kpid} weights the respective channel in the loss function. The falloff constant k_{kpid} will be helpful to define meaningful accuracy metrics in a later stage.

key point ID ($kpid$)	key point	key point loss weight ρ_{kpid}	falloff constant k_{kpid}
0	nose	1	0.052
1	right eye	1	0.5
2	left eye	1	0.05
3	right ear	1	0.07
4	left ear	1	0.07
5	right shoulder	1	0.158
6	left shoulder	1	0.158
7	right elbow	1.2	0.144
8	left elbow	1.2	0.144
9	right wrist	1.5	0.124
10	left wrist	1.5	0.124
11	right ASIS	1	0.214
12	left ASIS	1	0.214
13	right knee	1.2	0.174
14	left knee	1.2	0.174
15	right ankle	1.5	0.178
16	left ankle	1.5	0.178
17	upper sternum	1.5	0.158

(a) Illustration of missing spacial information in a 2D image; only overlayed, red pixels contain depth information due to camera limitations

(b) Illustration of recorded data as point cloud in 3d space. Pixels without spacial information are not visualized.

FIGURE 17. Visualization of underlying spacial information of data.

TABLE 8. Specifications of the dataset.

	dataset **X**
data size	313GB
number of participants	11
frame rate per second	3
image height	840
image width	480
annotated images	548

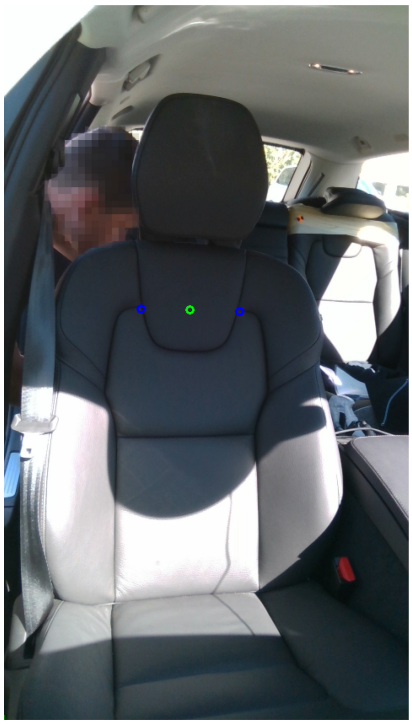


FIGURE 18. Visualization of point on seat, which has been used to normalize measurements of the torso movement. Blue dots have been selected by the distinctive shape of the seat, then the centroid (green) of them has been computed and the underlying spacial information have been used for further computations.

steps were done for all tps, resulting in the following Figure, which shows the frames in each cluster of the respective tp.

The training and validation dataset were obtained by randomly sampling from these clusters, resulting in around 340 images. 80 percent were used as training images and 20 percent of the images were used to validate the networks.

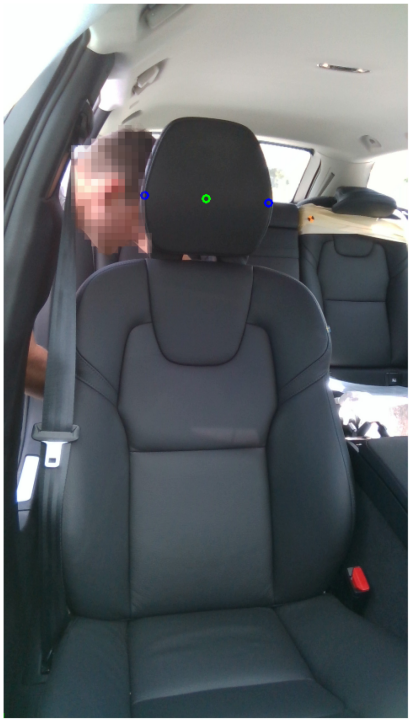


FIGURE 19. Visualization of point on seat, which has been used to normalize the measurement "Movement of the head in relation to the headrest." Blue dots have been selected by the distinctive shape of the head rest, then the centroid (green) of them has been computed and the underlying spacial information have been used for further computations.

TABLE 9. Information about study participants.

name	stature (m)	weight (kg)	gender (m/f)
tp8	1.72	n.a.	f
tp11	2.01	91	m
tp12	1.78	70	m
tp13	1.79	65	m
tp15	1.73	60	m
tp16	1.9	82	m
tp17	1.68	58	w
tp18	1.9	75	m
tp19	1.64	71	w
tp20	1.86	80	m

The heuristic approach described above ensured a high variance within the training and validation dataset.

In order to measure the average performance of the whole dataset, the test dataset has been sampled by equidistant sampling of every 500^{th} frame.

E. MEAN AVERAGE PRECISION

The mean average precision (for different thresholds ω), which is the single most important metric for the COCO key point detection task [9], was also provided for the present dataset in Table 10. However, it was not the single most important metric for the scope of this work, as a trade off towards a lower recall (by setting the confidence score of

(a) Side view.

(b) Front view.

FIGURE 20. Illustration of camera mount.

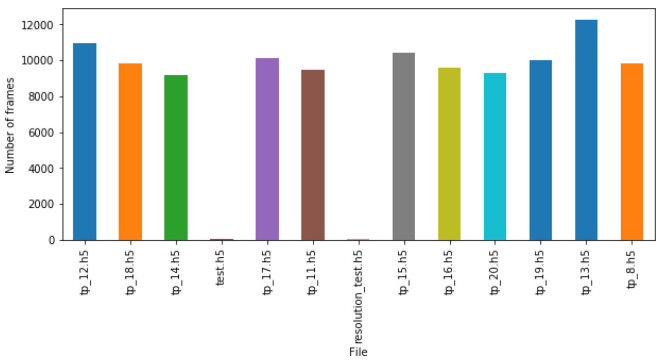


FIGURE 21. Frames per test person.

the region proposals of the 'Faster R-CNN ResNet-50 FPN' to 0.9) was made in order to obtain a high precision value.

F. TRAINING SEGMENTATION

The neural network was trained for 50 epochs with the ADAM optimizer and a weight decay regularization of 0.00003, the implementation of the weight decay regularization follows [25]. The learning rate for the ResNet-101 backbone encoder part was 0.001 and the learning rate for the FPN decoder part was 0.005. It was reduced by

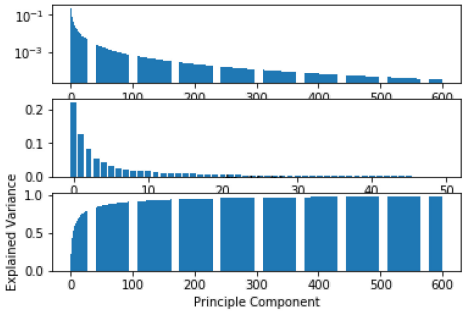


FIGURE 22. PCA components of tp16. The figure on top depicts the first 600 PCA components ordered by magnitude on a logarithmic scale. The middle figure depicts the first 50 PCA components and the bottom figure explains the accumulated variance by the PCA components.

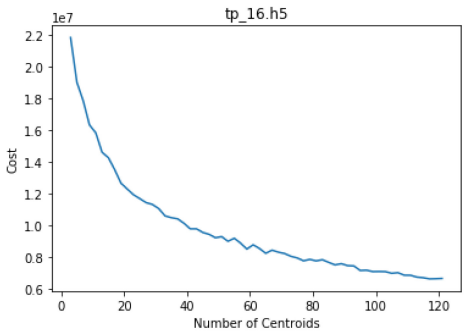


FIGURE 23. Illustration of the cost of a k-Means clustering algorithm for a varying number of centroids, performed on the PCA components, which explain 75 percent of the variance of tp 16.

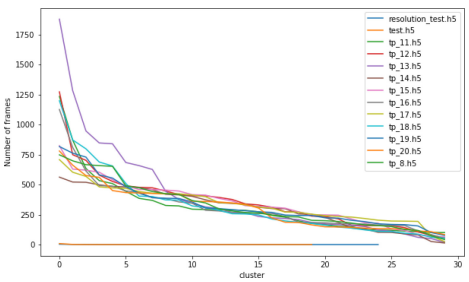


FIGURE 24. Number of frames per cluster for each tp.

TABLE 10. Performance of the HR-Net refined on the annotated testset X. $AP^{.50}$ and $AP^{.75}$ denote AP(X, 0.5, $N = 183$) and AP(X, 0.75, $N = 183$) respectively. AP^{M} and AP^{L} denote the metric *mAP* for different object scales. In the present case, there have only been bounding boxes exceeding an area of 9216 pixels, which is the threshold to count as a large (*L*) image. mAP is computed as a mean across the thresholds $\omega \in \{0.5, 0.55, 0.6, \ldots, 0.95\}$, interpolated across 100 recall values across all image sizes (only L present).

Method	mAP	$AP^{.50}$	$AP^{.75}$	AP^{M}	AP^{L}
HR-Net	0.731	0.802	0.802	-1.000	0.731
HR-Net refined	0.789	0.811	0.801	-1.000	0.789

a factor of 0.25, if the metric evaluated on the validation dataset was not improving for 3 epochs, a heuristic approach to get better generalization. Additional regularization was conducted by applying dropout on the concatenated FPN blocks with a rate of 0.25. The reason for different learning rates is due to the usage of pre-trained weights of the encoder part. These pre-trained weights were obtained by training the ResNet-101 architecture on the ImageNet classification task [26]. The pre-trained weights of the library Segmentation Models [27] were used. The was to utilize the pre-trained encoder part as a feature extractor, which is only slightly adapted to the data by a low learning rate, while accumulating features in the FPN decoder part, which could be used to segment the image.

The training dataset was augmented with the help of the library albumentations [28], which performed the following transformations on an image each time it was presented to the neural network:

1) A random rotation by 90, 180 or 270 degree with a probability p = 0.5.
2) A cutout of random image pixels: By a probability of 50 percent, 8 pixel locations were chosen. For each of these image locations, a width W and height H is uniformly selected from an integer value of 1 to 8 and then the rgb channels around the $W \times H$ area of the respective location were set to 0.
3) A change in contrast or brightness by a factor of 0.1 with a probability of 0.1
4) A grid distortion with probability 0.1
5) A change in hue, saturation or value of the image: With a probability of 0.5, an uniformly selected shift in hue within the interval $[-20, 20]$, a uniformly selected shift in saturation within the interval $[-30, 30]$ and a uniformly selected shift in rgb-channel value in the interval $[-20, 20]$ was applied.

The validation and test dataset were not augmented, because an optimal performance on undisturbed images was desired. The neural network were trained with a mini-batch size of 8 images, the training progress was measured with the metric *IoU* (dataset **X** *IoU* in Table 1) and trained with a weighted surrogate loss function consisting of Dice, Binary Cross Entropy (*BCE*) and *IoU* metrics (dataset *X WM* in Table 1) as an heuristic approach to obtain better class separation[1]:

$$\begin{aligned} \textit{weighted metrics } (WM) = 1.0 * \textit{Dice} \\ + 1.0 * BCE + 0.8 * IoU. \end{aligned}$$

Early stopping was applied, i.e., the best model state was picked according to the validation dataset and the metric *IoU* after each of the training processes.

1) RELATION OF PERSON DETECTION AND KEY POINT ACCURACY

To get a better perspective on the performance of the pipeline, precision and recall for different OKS-threshold levels ω for the whole dataset X with N = 222 samples were provided for the refined HR-Net:

The recall level of $\omega = 0.5$ in Table 11 can be explained by a large number of not detected bounding boxes. Loosely

1. *IoU*, *Dice* and *BCE* were defined in the Catalyst documentation [29].

TABLE 11. Precision and recall values for various OKS-threshold levels of the refined HR-Net.

Threshold	Precision	Recall
$\omega = 0.5$	0.9836	0.8219
$\omega = 0.75$	0.9726	0.8202
$\omega = 0.9$	0.9398	0.8151
$\omega = 0.95$	0.8743	0.8040

speaking, the recall level is the fraction of the correctly classified predictions (in terms of $\text{OKS}(\hat{bbp}_j) > \omega$) divided by the sum of the correctly classified predictions and the missing bounding boxes, which were not detected.

183 bounding boxes were actually detected by the 'Faster R-CNN ResNet-50 FPN'. 180 of them exceed the OKS classification threshold of $\omega = 0.5$. Thus the corresponding recall level is $180/219 = 0.8219$.

The other recall levels in Table 11 are roughly the same, which means that the bounding boxes, which were detected by the 'Faster R-CNN ResNet-50 FPN' [21], were a good basis for the consecutive key point prediction task, as the recall did not significantly drop.

The precision was the most important metric in the scope of this work, as it measured the fraction of the detected and correctly classified bounding boxes divided by all detected bounding boxes. It was crucial, that the key point predictions of the detected bounding boxes had a good quality, as these were the basis for further computations. This is the case, as the precision was around or above 0.95 for the recall levels $\omega \in \{0.5, 0.75, 0.9\}$. It dropped significantly for the threshold $\omega = 0.95$, however the lower level of $\omega = 0.9$ were considered good enough for the following computations.

ACKNOWLEDGMENT

The authors thank Ralf Werner, Koen Vallenga, Asli Pehlivan Rhodin, Herman Johnsson, and Tomas Björklund for their great support and helpful insights and discussions. This study was conducted in a collaboration of the Big Data Team and the Safety Center at Volvo Cars supported by Augsburg University.

REFERENCES

[1] "Euroncap Full Width Frontal Impact Testing Protocol, Version 1.2." EuroNCAP.com. 2019. [Online]. Available: https://cdn.euroncap.com/media/53141/euro-ncap-frontal-fw-test-protocol-v12.pdf (accessed Jun. 3, 2021).

[2] M. Reed, S. Ebert, M. Jones, and J. Hallman, "Prevalence of non-nominal seat positions and postures among front-seat passengers," *Traffic Injury Prevent.*, vol. 21, no. S1, pp. 7–12, 2020. [Online]. Available: https://doi.org/10.1080/15389588.2020.1793971

[3] "Euro NCAP 2025 Roadmap: In Pursuit of Vision Zero." EuroNCAP.com. 2017. [Online]. Available: https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf (accessed Apr. 26, 2021)

[4] E. Marcy. "Iihs Research: Virtual Testing for Out-of-Position Occupants in Low Severity Rear Impacts." 2020. [Online]. Available: https://projectvirtual.eu/wp-content/uploads/2020/09/07-Marcy_Virtual-Workshop-Presentation-IIHS-Edwards-2020.pptx (accessed Apr. 26, 2021).

[5] E. Zadobrischi, L.-M. Cosovanu, and M. Dimian, "Benefits of a portable warning system adaptable to vehicles dedicated for seat belts detection," in *Proc. 24th Int. Conf. Syst. Theory Control Comput. (ICSTCC)*, 2020, pp. 892–897.

[6] A. Ş. Şener, I. F. Ince, H. B. Baydargil, I. Garip, and O. Ozturk, "Deep learning based automatic vertical height adjustment of incorrectly fastened seat belts for driver and passenger safety in fleet vehicles," *Proc. Inst. Mech. Eng. D, J. Automobile Eng.*, Jun. 2021, Art. no. 095440702110253. [Online]. Available: https://doi.org/10.1177/09544070211025338. doi: 10.1177/09544070211025338.

[7] Q. Yi and Q. Yi, "Safety belt wearing detection algorithm based on human joint points," in *Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, pp. 538–541, 2021.

[8] M. Baltaxe, R. Mergui, K. Nistel, and G. Kamhi, "Marker-less vision-based detection of improper seat belt routing," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 783–789. [Online]. Available: http://dblp.uni-trier.de/db/conf/ivs/ivs2019.html#BaltaxeMNK19

[9] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision (ECCV)*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 740–755. [Online]. Available: http://cocodataset.org/

[10] S. Chun *et al.*, "NADS-Net: A nimble architecture for driver and seat belt detection via convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2019, pp. 2413–2421.

[11] X. Zeng, S. Bei, E. Wang, W. Luo, and T. Liu, "A method of learner's sitting posture recognition based on depth image," in *Advances in Intelligent Systems Research*, vol. 134. Paris, France: Atlantis Press, Jan. 2017, pp. 558–563.

[12] C.-C. Wu, C.-C. Chiu, and C.-Y. Yeh, "Development of wearable posture monitoring system for dynamic assessment of sitting posture," *Phys. Eng. Sci. Med.*, vol. 43, pp. 187–203, Dec. 2019.

[13] J. Wang, B. Hafidh, H. Dong, and A. E. Saddik, "Sitting posture recognition using a spiking neural network," *IEEE Sensors J.*, vol. 21, no. 2, pp. 1779–1786, Jan. 2021.

[14] W. Sun, Z. Zhou, and H. Li, "Sitting posture recognition in real-time combined with index map and BLS," in *Proc. ICIAI*, 2019, pp. 101–105. [Online]. Available: https://doi.org/10.1145/3319921.3319955

[15] M. P. Reed, S. M. Ebert, B.-K. D. Park, and M. L. H. Jones, "Passenger kinematics during crash avoidance maneuvers," Transport. Res. Inst., Univ. Michigan, Ann Arbor, MI, USA, Rep. UMTRI-2018-5, 2018.

[16] Y. Xing *et al.*, "Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 95–108, Mar. 2018.

[17] H. Wang, M. Zhao, G. Beurier, and X. Wang, "Automobile driver posture monitoring systems: A review," *Zhongguo Gonglu Xuebao/China J. Highway Transp.*, vol. 32, pp. 1–18, Feb. 2019.

[18] *Intel Realsense Product Family D400 Series Datasheet*, document 337029-009, Intel RealSense, Santa Clara, CA, USA, 2020. Accessed: Jan. 18, 2021. [Online]. Available: https://www.intelrealsense.com/depth-camera-d415/

[19] J. Brooks. "COCO Annotator." GitHub.com. 2019. [Online]. Available: https://github.com/jsbroks/coco-annotator/ (accessed Jan. 18, 2021).

[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2016, *arXiv:1506.01497*.

[22] *CloudCompare, Version 1.19*. (2020). Open Source Software. Accessed: Jan. 18, 2021. [Online]. Available: http://www.cloudcompare.org/

[23] C. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[24] V. Guruswami and R. Kannan (Carnegie Mellon Univ., Pittsburgh, PA, USA). *Computer Science Theory for the Information Age—Chapter 'Singular Value Decomposition*. (2012). Accessed: Jan. 18, 2021. [Online]. Available: https://www.cs.cmu.edu/ venkatg/teaching/CStheory-infoage/

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019, *arXiv:1711.05101*.

[26] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[27] P. Yakubovskiy. "Segmentation Models." GitHub.com. 2019. [Online]. Available: https://github.com/qubvel/segmentation_models

[28] A. Buslaev, A. Parinov, E. Khvedchenya, V. Iglovikov, and A. Kalinin, "Albumentations: Fast and flexible image augmentations," 2018, *arXiv:1809.06839*.

[29] S. Kolesnikov. "Accelerated Deep Learning Research and Development." GitHub.com. 2018. [Online]. Available: https://github.com/catalyst-team/catalyst (accessed Jan. 18, 2021).